Development and Application of Tools for MRI Analysis - A Study on the Effects of Exercise in Patients with Alzheimer's Disease and Generative Models for Bias Field Correction in MR Brain Imaging

Larsen, Christian Thode

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Development and Application of Tools for MRI Analysis

A Study on the Effects of Exercise in Patients with Alzheimer's Disease and Generative Models for Bias Field Correction in MR Brain Imaging

Christian Thode Larsen

Kongens Lyngby 2015
PHD-2015-378

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Richard Petersens Plads, building 324,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk

PHD: ISSN 0909-3192

Summary

Magnetic resonance imaging (MRI) is the *de facto* modality in neuroimaging studies, due to its superior image contrast in soft tissue. These studies often employ automated software pipelines that segments the image into structures and tissue. This reduces the time needed for analysis as well as statistical bias that may arise due to disagreements in delineations made by human experts. One such pipeline is Freesurfer.

This thesis presents results from the intervention study "Preserving cognition, quality of life, physical health and functional ability in Alzheimer's disease: the effect of physical exercise" (ADEX), where longitudinal Freesurfer analysis was used to obtain segmentations of the hippocampal subfields and cortical regions in a subgroup of participants before and after a four-month exercise period. The participants performed moderate-to-high aerobic exercise for one hour, three times per week. The study hypothesized that the intervention would lead to reduced loss of tissue in the hippocampus and cortical regions, and that volumetric changes over time would correlate with cognitive performance measures. It was not possible to measure any effects in the hippocampus or cortical regions due to the intervention. However, it was found that exercise load (attendance and training intensity) correlated with changes in the hippocampus and in frontal and cingulate cortical thickness. Furthermore, changes in frontal and cingulate cortical thickness were found to correlate with changes in several cognitive performance measures, including mental speed, attention and verbal fluency.

MRI suffers from an image artifact often referred to as the "bias field". This effect complicates automatized analysis of the images. For this reason, bias field

correction is typical an early preprocessing step in many pipelines. Freesurfer currently employs the popular N3 bias field correction algorithm early in the pipeline, to solve this problem.

In this thesis, the reader is introduced to generative models for bias field correction. It is further shown how N3, which has traditionally been described as a "histogram sharpening" method, actually employs an underlying generative model, and that the bias field is estimated using an algorithm that is identical to generalized expectation maximization, but relies on heuristic parameter updates. The thesis progresses to present a new generative model for longitudinal correction of the bias field, as well as a model that does not require brain masking or probabilistic, anatomical atlases in order to perform well. Finally, the thesis presents the realization of these models in the software package "Intensity Inhomogeneity Correction", which will be made publicly available.

Resume

Magnetic resonance imaging (MRI) er den dominante billedmodalitet i neuroimaging studier givet dens overlegne billedkontrast i blødt væv. Disse studier benytter ofte automatiserede software pipelines som segmenterer billedet i strukturer og væv. Dette reducerer det nødvendige tidsforbrug i analysen såvel som statistisk bias der måtte opstå på grund af uoverensstemmelser i manuelt indtegnede segmenteringer, lavet af menneskelige eksperter. Én sådan pipeline er Freesurfer.

Denne afhandling omhandler delresultater fra interventionsstudiet "Effekten af fysisk træning på livskvalitet, fysisk helbred og funktionsevne hos patienter med Alzheimers sygdom" (ADEX). Longitudinel Freesurfer analyse blev anvendt til at lave segmenteringer af hippocampus samt kortikale regioner i en undergruppe af deltagere, før og efter en træningsperiode på 4 måneder. Deltagerne udførte mellem-til-høj intensitet aerob træning 3 gange ugentligt af en times varighed. Studiets hypotese var, at interventionen ville medføre mindre vævstab i hippocampus og kortikale regioner, samt at volumetriske ændringer over tid ville korrelere med ændringer i kognitive mål. Det var ikke muligt at påvise effekt af interventionen på hippocampus eller kortikale volumenmål. Motionsmængde (fremmøde samt intensitet) blev fundet at korrelere med ændringer i hippocampus volumen samt med tykkelse af frontal kortex og gyrus cingularis. Ydermere korrelerede ændringer i frontal kortex og gyrus cingularis tykkelse med ændringer i forskellige kognitive mål, herunder mental hastighed, opmærksomhed og ordmobilisering.

MRI lider under en billedartifakt ofte kaldet "biasfeltet". Denne artifakt komplicerer automatiseret analyse af billederne. Af denne grund er bias felt korrektion ofte et tidligt preprocesserings skridt i mange pipelines. Freesurfer benytter

på nuværende tidspunkt den populære N3 bias felt korrektionsmetode tidligt i sin pipeline for at løse dette problem.

I denne afhanding introduceres læseren til generative modeller af bias felt korrektion. Det vises yderligere hvordan N3, som traditionelt er blevet beskrevet som en "histogram skærpende" metode, faktisk bygger på en underliggende generativ model, og at bias feltet estimeres med en algoritme som er lig generaliseret expectation maximization, men hvor der bruges heuristiske parameter opdateringer. Afhandlingen fortsætter med at præsentere en ny generativ model af longitudinal korrektion af bias feltet, såvel som en model der ikke er afhængig af hjerne masker eller probabilistiske, anatomiske atlaser for at opnå gode resultater. Endeligt præsenterer afhandlingen realiseringen af disse modeller i software pakken "Intensity Inhomogeneity Correction", som vil blive gjort offentligt tilgængelig.

Preface

In 2011 I finished my master studies. At this point in time I had obtained a highly specialized background within computer graphics, physically-based rending of images and software engineering. With no medical background and only very limited proficiency within mathematical modeling and multivariate statistics, it may therefore seem odd that I chose a PhD study that would bring me to work within these areas. In particular, fields of research covering Bayesian modeling, neuroimaging and clinical MRI studies. However, an interest in learning new things drove me forward. The study proved a challenge, but it was ultimately a mountain that I was able to climb.

Kongens Lyngby, Denmark, August 2015

Christian Thode Larsen

Acknowledgements

Acknowledgements are warranted for several people, each of which either collaborated, helped or otherwise supported me and my work throughout this PhD study.

Main supervisor Koen Van Leemput was invaluable during the process of becoming sufficiently proficient with generative modeling of bias field correction, in order to complete the work done throughout this PhD. Without his extensive insight into this field of research, I would not have come as far as I have. Furthermore, Koen made it possible for me to spend six months collaborating at the Laboratory for Computational Neuroimaging (LCN), Athinoula A. Martinos Center for Biomedical Imaging, Boston, USA. This stay was in many ways invaluable to me, both professionally as well as personally.

Also from DTU, Oula Puonti was a big help with discussions of generative modeling theory and technical details. Anders Nymark Christensen assisted with multivariate statistics in the ADEX project. This was fundamental for reaching a successful statistical analysis and conclusion of the study. Finally, Rasmus Larsen, Knut Conradsen and Anders Dahl all helped faciliate my study, and all of my colleagues at DTU made the years a pleasant experience.

Co-supervisor Ellen Garde was very helpful in facilitating my work at the Danish Research Center for Magnetic Resonance (DRCMR) and in the ADEX study. Technical and clinical research are quite different, and Ellen was a tremendous help in the early process of bringing me into an entirely new environment.

Hanne Schmidt and Sascha Gude from the reader center and support group at

DRCMR did manual work that was required in order to successfully analyze the MRI data from the ADEX study. Without their huge effort, the extent of the analysis would have been impossible to overcome. Hanne and Sascha, together with many other people from DRCMR, made my time there a pleasant experience. Sussi Larsen should also be acknowledged for her hard work in obtaining MRI scans of the ADEX study participants, a task that was by no means easy. Finally, both Kristian Steen Frederiksen and Steen Hasselbalch from the Danish Dementia Research Centre, Rigshospitalet, Denmark together with Hartwig Siebner from DRCMR were all supportive of my work and very helpful in facilitating the collaboration between hospitals and DTU. Kristian was kind to provide comments and suggestions for this thesis.

From the Martinos Center in Boston, several people deserve thanks. J. Eugenio Iglesias in particular was a great help in my work on generative modeling of bias field correction, and together with the rest of group from the Laboratory for Computational Neuroimaging made my stay quite enjoyable. Jonathan Polimeni provided 7T test data for my work, and both Douglas Greve and Martin Reuter were helpful with assistance on how to use the Freesurfer pipeline, both during the my stay in Boston, and after. Jon, Doug and Martin were all kind to provide comments and suggestions for this thesis.

Both family and friends naturally deserve many thanks. They all supported me and my decision to do this PhD study throughout several, very busy years. Particularly I thank Bruce Clarke and Barry Laterman whom I lived with while in Boston. They were a great support during a very busy and challenging period of time.

Finally, Sarah Risinger deserves a very special thank you, as she showed me the world is made of much more than simply mountains that need to be climbed.

Scientific Contributions

Papers Included in this Thesis

- Larsen, C. T., Frederiksen, K. S., Hasselbalch, S. G., Christensen, A. N., Høgh, P., Wermuth, L., Lolk, A., Andersen, B. B., Siebner, H., Waldemar, G. and Garde, E. 2015. Effect of moderate-to-high intensity aerobic exercise on hippocampus and cortical regions in patients with mild to moderate Alzheimer's disease. Journal manuscript, title tentative.

- Larsen, C. T., Iglesias, J. and Van Leemput, K. 2014. N3 bias field correction explained as a bayesian modeling method. In Bayesian and grAphical Models for Biomedical Imaging, vol. 8677 of Lecture Notes in Computer Science. Springer International Publishing, 1-12.

  Best paper award, Bayesian and Graphical Models for Biomedical Imaging (BAMBI) workshop at the 17th International conference on Medical Image Computing and Computer Assisted Interventions (MICCAI) 2014.

- Larsen, C. T., Iglesias, J. E. and Van Leemput, K. 2015. A unified, generative model for bias field correction of cross-sectional MRI: re-evaluating and alleviating the need for brainmasking and anatomical atlases. Journal manuscript, title tentative.

Other Papers

- Larsen, C. T., Frisvad, J. R., Jensen, P. D. and Bærentzen, J. A. 2012. Real-time rendering of teeth with no preprocessing. In Advances in Visual Computing, Vol. 7432 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, 334-345.

- Larsen, C. T., Gross, S. and Bærentzen, J.A. 2015. Removing lectures in a computer programming course – a quantitative study. Conference article, 11th International CDIO Conference, Chengdu, China.

- Christensen, A. N., Larsen, C. T., Dahl, V. A., Petersen, M. B. and Conradsen, K. 2015. Software-package for automated bias field correction and quantification of abdominal fat using MRI with preliminary evaluation in overweight subjects. Journal manuscript under preparation, title tentative.

Contents

Chapter 1

Introduction

This thesis serves as an introduction into a number of topics which were all central during the performed PhD study. The introduction provides the reader a) some insight into the background and theory underlying the three, included papers, should he not already be familiar with the topics, and b) an overview of the theory that the author had to familiarize himself with throughout. The topics covered are as a whole a somewhat interesting mix, hence the reason for the thesis title, which aims to capture how it was necessary to "build a bridge" between two seemingly different areas of research: one focusing on practical application of tools for image analysis, and the other on their development.

At the center of the study is magnetic resonance imaging (MRI), as the analysis of images produced by this technique was explored both from a practical as well as a theoretical point view. As such, chapter 2 provides a basic insight into MRI, including the underlying theoretical basics for the technique, a description of how a broad range of images can be acquired using the technique and finally some technical challenges associated with the acquisition.

One part of the PhD study was dedicated to longitudinal analysis of MR brain images acquired under one branch of the national research project "Preserving cognition, quality of life, physical health and functional ability in Alzheimer's disease: the effect of physical exercise" (ADEX). This analysis was performed using a fairly elaborate processing pipeline "Freesurfer". The work under this

project were carried out at the Danish Research Center for Magnetic Resonance (DRCMR) with co-supervisor Ellen Garde, MD. PhD. and project leader and professor Hartwig Siebner, MD. PhD. Furthermore, the ADEX study involved many collaborators, in particular the Memory Disorders Research Group, Danish Dementia Research Center, Department of Neurology, Copenhagen as well as the Department of Applied Mathematics and Computer Science, Technical University of Denmark (DTU). This work is covered in chapter 3 which provides an introduction to paper A.

The other part of the PhD study was dedicated to the development of methods for bias field correction, employing Bayesian (generative) models. Bias field correction is a very important ingredient in most analyses of MRI, and it is one of the first processing steps in the Freesurfer pipeline. This work was carried out at DTU with main supervisor and associate professor Koen Van Leemput, PhD. and in collaboration with the Laboratory for Computational Neuroimaging at the Athinoula Martinos Center, Massachusetts General Hospital, Harvard University, Boston, USA[1] under the supervision of J. Eugenio Iglesias. This work is covered in chapter 5, and provides an introduction to papers B and C.

The glue that binds these topics together then is the analysis itself: in order to *motivate* it you need a concrete study, in this case the effect of exercise in patients with Alzheimer's disease. And to *realize* it you need the right toolbox, in this case Freesurfer and bias field correction. This means that practical studies rely on the application of usable tools for analysis, which motivates their development. Conversely, by developing the tools we improve the toolbox, which may allow new studies to be done which were previously not technically possible, or old studies to be revisited because the tools became better and more accurate. As such, these two topics are intertwined (development and application), and one cannot (reasonably) exist without the other.

In the case of this study, theoretical as well as practical development of the toolbox for image analysis was the realization of computational models for bias field correction in a software package named "Intensity Inhomogeneity Correction" (IIC).

Chapters 6 and 7 concludes this thesis by discussing potential for future work, as well as observations made throughout the study that was not previously covered in other chapters.

[1] Including six months of collaboration on-site.

Chapter 2

Magnetic Resonance Imaging

This chapter provides the reader with a brief introduction to MRI. Section 2.1 discusses the working principles behind MRI, and Section 2.2 how the technology can be used to generate images of very different properties. Given that the focus of this thesis is analysis of MR images, technicalities such as elaborate MR physics, details of MR pulse sequences and associated hardware are not covered. Section 2.1 and 2.2 are generally based on relevant explanations of MRI given in the note "Introduction to Magnetic Resonance Imaging Techniques" by Lars G. Hanson. The final Section 2.3 of this chapter describes an artifact inherent in all MR images, which typically needs to be addressed in some way in order for automated methods to process the MR data successfully.

2.1 Working Principles

We start this introduction to MRI by considering the main, static magnetic field in the MR scanner. We refer to this field henceforth as the $\boldsymbol{B}_0$ *field*. The strength of the field (magnitude of $\boldsymbol{B}_0$) has unit *Tesla* (T). The strength of the $\boldsymbol{B}_0$ field is commonly referred to simply as the "field strength", e.g., 3T or 7T on modern scanners. Figure 2.1 shows a modern 7T scanner from Siemens.[1]

[1] Illustration from `http://www.healthcare.siemens.com`

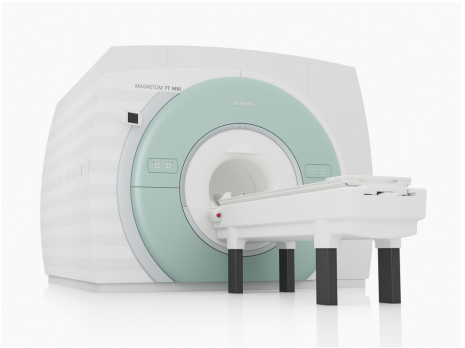


Figure 2.1: Siemens Magnetom 7T MRI Scanner.

The bearing principle in MRI is that hydrogen nuclei (protons) possess spin (rotation), which makes them magnetic along the axis they spin around. When the nuclei are placed within the $\boldsymbol{B}_0$ field, they will precess around it. Precession is analogous to the rotation of a pendulum around a vertical axis. The precession frequency of the nuclei is given by the Larmor equation

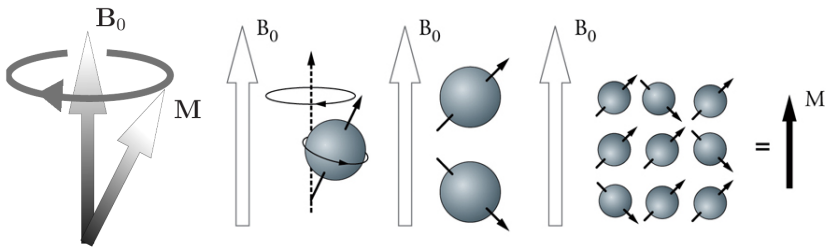
$$f = \gamma \boldsymbol{B}_0, \tag{2.1}$$

where γ is a constant ($42MHz/T$ for hydrogen protons). The *net-magnetization* $\boldsymbol{M}$ of the nuclei similarly precess around the $\boldsymbol{B}_0$ field until it is aligned. We refer to this stage as *equilibrium* and the process as *relaxation*. The net-magnetization moves towards equilibrium due to interactions between nuclei at near-collisions. A simple diagram illustrating precession has been shown in Figure 2.2.[2]

While the nuclei precess, they emit radio waves (electromagnetic fields that change in time) with the same frequency as the precession. We refer to this as the *resonance* frequency. As governed by the Larmor equation, the higher the field strength of the scanner is, the higher the frequency and consequent signal. At full relaxation, the nuclei no longer emit radio waves. Consequently, to change, preserve or restore precession of the nuclei in order to measure a signal (continuously), they must be pushed out of equilibrium again. This is done by applying a second magnetic field by means of transmission coils. We refer to this field as the $\boldsymbol{B}_1$ *field* and the process as *excitation*.

Excitation can be considered intuitively by using a compass as an example: when the needle (comparable to a single hydrogen proton) has aligned with the north-direction (the earth's magnetic field), it is possible to "push" the needle by placing a magnet close to it. When removed from its vicinity, the needle will

[2]Illustrations from "Introduction to Magnetic Resonance Imaging Techniques, Lars G. Hanson (a) and `http://i.stack.imgur.com/wJtmZ.jpg` (b)



(a) (b)

Figure 2.2: a) First illustration: the magnetization $\boldsymbol{M}$ of a single nucleus that precesses around the magnetic, static field $\boldsymbol{B}_0$. b) Three illustrations showing how the precession of several nuclei leads to a net magnetization $\boldsymbol{M}$ that aligns with the $\boldsymbol{B}_0$ field at equilibrium.

start to align with the earths magnetic field again. By repeating this process, it is possible to make the needle oscillate around the north direction. Another example is that of pushing a person on a swing. By applying energy (pushing) with the same frequency as the swing, it will swing harder. Conversely, if the energy is not applied with the same frequency, the effect will be much less effective. In this example, gravity would correspond to the $\boldsymbol{B}_0$ field.

2.2 Obtaining Images

The magnetic field emitted during the precession of the nuclei, given their net-magnetization $\boldsymbol{M}$, can be measured with proper equipment, such as the receive coils in the MR scanner. Different types of tissue, e.g., white matter (WM), gray matter (GM) or cerebrospinal fluid (CSF) in the brain, have different consistency, which affects how freely the hydrogen nuclei are allowed to move. As such, the time it takes for the net-magnetization in each tissue to relax differ. The relaxation of the nuclei for a given tissue can be considered in terms of longitudinal magnetization $\boldsymbol{M}_z$, and transversal magnetization $\boldsymbol{M}_{xy}$ in a plane perpendicular to the longitudinal magnetization. Each of these relax on different time scales. $\boldsymbol{M}_z$ relaxes linearly on a timescale referred to as *T1*. The transversal magnetization $\boldsymbol{M}_{xy}$ relaxes exponentially until it reaches $\boldsymbol{M}_0$ on a timescale referred to as *T2*.

Applying a pulse sequence (a collection of RF and gradient pulses) is how the $\boldsymbol{B}_1$ field is created in order to excite the nuclei. Each pulse sequence is composed by a number of RF and gradient pulses lasting for some duration, in a particular

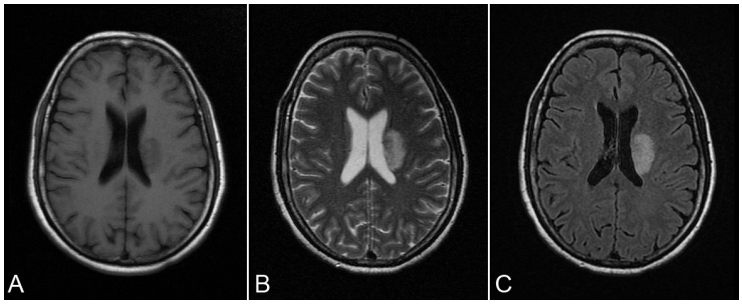


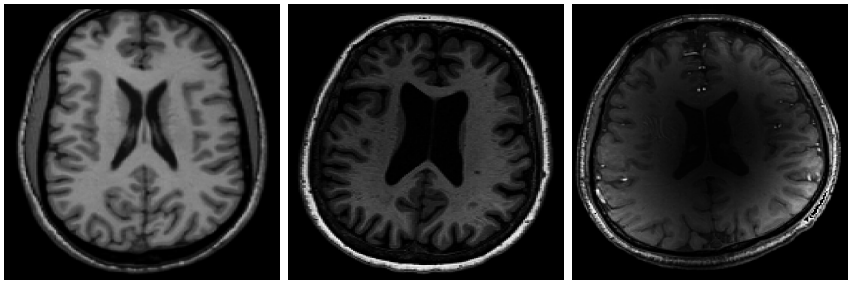
Figure 2.3: Images obtained using A) T1-weighed, B) T2-weighed and C) FLAIR sequences respectively. White matter appear bright in A) but dark in B) and somewhere in between in C). Conversely, CSF appears dark in A), very bright in B) and has been fully suppressed in C). The hyperintense blob in B and C (hypointense in A) is due to a stroke.

order and with some time between them. The composition of the pulse sequence consequently allows one to manipulate how the nuclei relax along the transversal and longitudinal axis, i.e., how the different time scales in the image are weighed. Given the many ways a pulse sequence can be configured, a broad range of images with intensity and contrast properties can be created, such as

- *T1-weighed* images that has enhanced contrast with respect to different tissue types, with WM being very bright, GM intermediary and CSF dark.
- *T2-weighed* images where the intensity of fluid appears very bright and tissue is relatively dark.
- *FLuid-Attenuated Inversion Recovery (FLAIR)* images has fluid nulled and preserves signal in WM and GM, resulting in intensities falling somewhere between bright and dark.

Figure 2.3[3] illustrates a T1-weighed, a T2-weighed and FLAIR image respectively.

[3] `http://www.biomedcentral.com/content/figures/1471-2377-11-49-2-l.jpg`



(a) (b) (c)

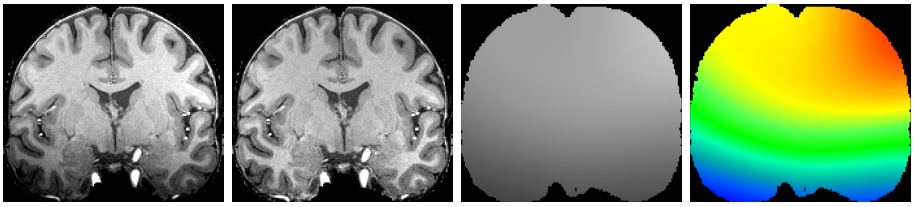
Figure 2.4: The difference between the effect of the bias field on images recorded at (a) 1.5T (b) 3T and (c) 7T respectively.

2.3 The Bias Field Artifact

MRI suffers from a particular imaging artifact commonly referred to as "intensity inhomogeneity" or "bias field", which appears as a low-frequent "noise" in the image. More specifically, the voxels in the image appear brighter or darker than they're supposed to, an effect that sometimes resembles that of a flashlight. While MRI is actually affected by many bias effects, here we consider only those that create this "intensity" bias, which is generally assumed to be multiplicative within a tissue. This assumption is made in most bias field correction methods in order to model the bias field effect, which will be elaborated upon in chapter 5.

The bias field artifact is present at all magnetic field strengths, and is caused by inhomogeneities in transmit field efficiency and receive field sensitivity. As the field strength increases (e.g., 7T), so does the effects due to transmit field efficiency which is dictated by the object being scanned, specifically its shape, position and orientation, and more generally its permeability and dielectric properties (the degree to which the spins can be excited). Conversely, the effects due to the receive field depend more on the (array of) coils being used for reception. Figure 2.4 illustrates the difference between the effect of the bias field on images recorded using a 1.5T, 3T and 7T scanner respectively.

Since intensity inhomogeneity negatively impacts any computerized analysis of MRI data, its correction is highly important. Effects that cannot be corrected using shimming techniques[Liang and Lauterbur 2000; Chen et al. 2004], need to be corrected post image acquisition, and methods for doing this are often applied as one of the first steps in MRI analysis pipelines. One example is the popular N3 algorithm [Sled et al. 1998] employed in Freesurfer [Fischl et al.



(a) (b) (c) (d)

Figure 2.5: An illustration of the bias field artifact in a 7T image. From left to right: a) the uncorrected data, b) the corrected data using the popular N3 bias field correction algorithm, c) estimate of the bias field in gray scale and d) with heap map coloring.. White matter and gray matter tissues appear much more homogeneous in the corrected image, with notable improvements in the temporal lopes, in particular.

2002]. Figure 2.5 shows the effect of the bias field on an image obtained on a 7T scanner, an estimate of the bias field using the popular N3 algorithm and the corresponding corrected image.

It should be noted that the inhomogeneity resulting from the transmit field is not multiplicative, although the effects of modeling it as such may be negligible at field strengths of 1.5T and even 3T, at least for the purpose of segmentation [Styner and Van Leemput 2004]. MR imaging are affected by many other artifacts, all of which are discussed in e.g., [Vovk et al. 2007]. More relevant causes and underlying physics for the bias field effect we are concerned with here, are discussed in e.g., [Collins et al. 2005]. Furthermore, how the effect worsens at field strengths of 7T and above are discussed in e.g., [de Moortele et al. 2005; Wrede et al. 2012].

A Note on Contrast Bias

Inhomogeneities in the transmit field can also cause "contrast" bias, in particular at high field strengths $\geq$ 7T, which makes e.g., white and gray matter voxels in brain MRI isointense (voxel intensities appear similar). Contrast bias is a problem which cannot be corrected using a multiplicative bias field model, but can to some extent be solved by using a proper MR sequence for image acquisition. The problem is discussed in greater detail in [Fujimoto et al. 2014]. As such, an interesting challenge in terms of collaboration between MR physicists and software engineers is to "solve" the bias field problem by devising pulse sequences that minimize contrast bias, and then correcting for the intensity bias post image acquisition.

Chapter 3

The ADEX MRI Study and Freesurfer

This chapter describes the ADEX study, its motivation, design and findings, as well as the tools used to successfully analyze the ADEX MRI data. Particular emphasis is put on how configuration and use of these tools may have significant consequences for study outcomes.

Section 3.1 provides the reader with a brief introduction to AD, including its symptoms, progression, effects on patient, family and caretakers and finally a current theory on its cause (which is unknown). Section 3.2 presents the study "Preserving cognition, quality of life, physical health and functional ability in Alzheimer's disease: the effect of physical exercise" (ADEX). Most details covered, such as the selection and screening of patients, randomization procedures and cognitive and physical testing were not executed in the MRI part of the study (which is the focus of this thesis). Still, some elaboration is necessary to present a more complete overview of ADEX and the associated MRI analyses. In Section 3.3 the brain analysis pipeline Freesurfer, which was used to analyze the ADEX MRI data, is described. Section 3.4 presents the contributions made in this thesis within the ADEX MRI study, and Section 3.5 more generally discusses the study, the data analysis and findings, and other methods of analysis that could be of interest. The final section describes and exemplifies the importance of proper bias field correction. This serves to motivate the following chapter 5 on generative models for bias field correction.

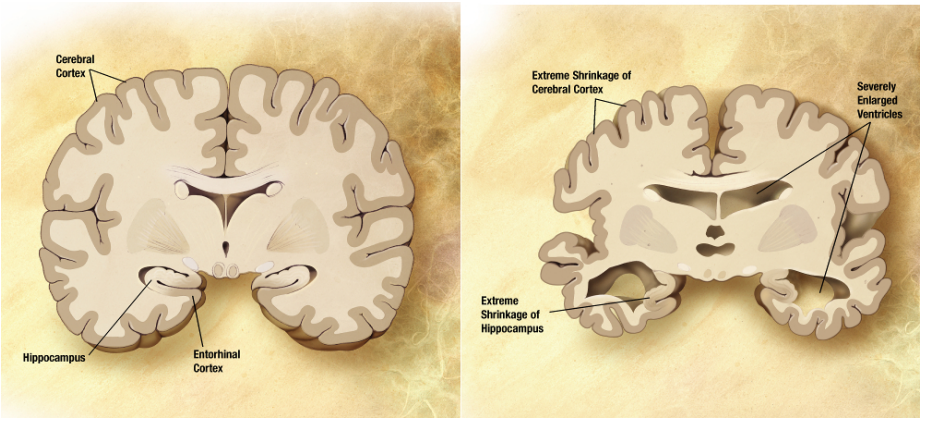


Figure 3.1: A comparison between a healthy brain (left) and a diseased AD brain (right). Key identifying features are a shrinking of the hippocampus, severe thinning of the cerebral cortex and enlarged ventricles.

3.1 Alzheimer's Disease

AD is one type of *dementia* (cognitive decline over time, leading to impaired activities in daily life). It is a neuro-degenerative disease, which leads to progressive *atrophy* in the brain (tissue "wasting away" due to cell degeneration). MR imaging studies have shown that atrophy can be observed in the anterior hippocampus already at a diagnosis of mild cognitive impairment (MCI), several years prior to full AD diagnosis. As the disease progresses, shrinkage of the hippocampus becomes more severe, and atrophy can be observed in other regions of the brain, in particular the temporal and parietal lobes at full AD diagnosis[Whitwell et al. 2007]. Other visual characteristics of this brain degeneration are general cortical thinning and enlarged ventricles. Figure 3.1 shows the main differences between a healthy and AD affected brain.[1]

The hippocampus plays an important role in short and long term memory, and also spatial orientation. Given the early involvement of the hippocampus in AD, these functions are progressively impaired over time, and in particular loss of episodic memory is one of the symptoms first observed [Burns and Iliffe 2009]. As AD progresses, other symptoms may start to emerge, such as rapid mood changes, behavioral and motivational issues and a failure to take proper care of one-self. All of these factors contribute to make activities in daily life very difficult for the AD patient. Furthermore, great strain is put on the people

[1] `https://upload.wikimedia.org/wikipedia/commons/a/a5/Alzheimer's_disease_brain_comparison`

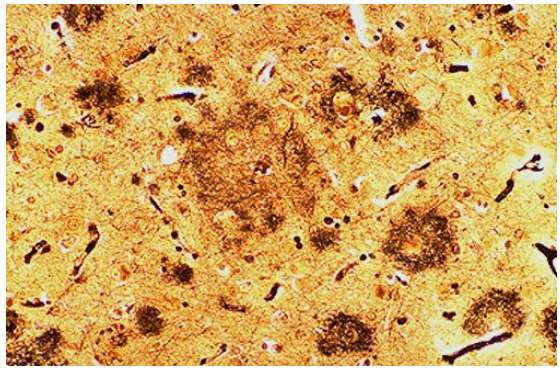


Figure 3.2: An illustration of amyloid plague in the brain of a patient with AD.

closest to the patient, including family, friends and caregivers.

The cause of AD is not well understood, and it has no cure. Current pharmacological treatments of AD exist, but they are symptomatic and at best may slow the rate of decline[Castellani and Perry 2012]. A number of theories for the cause of AD have been presented over time. The prevailing theory involves beta-amyloid (A_β) peptides from the amyloid precursor protein (APP) as a primary cause of neuro-denegeration, due to their involvement in amyloid plagues [Hardy and Allsop 1991], which can be observed in AD on a microscopic level. An illustration of an amyloid plague has been shown in Figure 3.2[2]. The theory on the involvement of A_β was updated in 2009 to suggest that the neuro-degenerative effects are due to a pruning of neuronal brain connections being triggered by processes related to aging [Nikolaev et al. 2009]. This process is normally seen in early life while the brain rapidly grows and is controlled by another component of APP (N-APP). Here, A_β only plays a complementary role, where the deposition of A_β may be a triggering factor, whereas other changes drives the neuro-degeneration. This theory only describes one of several possible mechanisms; the actual process of neuro-degeneration may be different or more involved.

Today (2015), around 30 million people have been diagnosed with AD, which constitutes 60-70% of the patients diagnosed with dementia.[3] This makes AD very costly to society, and together with the lack of a cure and pharmacological treatment, these factors illustrate why studies into finding alternative ways of improving activities of daily life in patients with AD, at least for a while, are so important. Exercise has been suggested as one such way.

[2] `http://library.med.utah.edu/WebPath/jpeg5/CNS090.jpg`.

[3] `http://www.who.int/mediacentre/factsheets/fs362/en/` (Dementia Fact Sheet N°362, World Health Organization).

3.2 The Effects of Exercise in AD

The ADEX study is to our knowledge the first of its kind to explore the effects of moderate-to-high aerobic exercise in patients with mild-to-moderate AD. ADEX was motivated by recent studies on the effects of exercise in healthy elderly, which suggests effects such as improved cognition [Vreugdenhil et al. 2012; de Andrade et al. 2013] and stimulated brain growth [Colcombe et al. 2006], including increases in hippocampal volume [Erickson et al. 2011] and pre-frontal and cingulate cortical volume [Ruscheweyh et al. 2011].

For this purpose, approximately two-hundred patients with mild AD were recruited from eight memory clinics[4] in Denmark. A non-exhaustive list of inclusion criteria was that the participants had a Mini Mental State Examination (MMSE) score of at least 20, were 50-90 years of age, had regular caregiving and were in general good health with functional sight and hearing. Conversely, risk factors that could potentially complicate physical activity, such as neurological, medical or psychiatric diseases, or alternatively an already high level of aerobic exercise, all lead to exclusion from the study.

The recruited participants were then randomized into a control and intervention group following a single-blind, randomized procedure. Cognition of all participants was assessed using a number of tests at baseline:

- "Symbol Digits Modalities Test" (SDMT): testing mental speed and attention,
- verbal fluency: assessing the number of words (semantic or phonetic) said from a given category over a fixed time interval,
- "Alzheimer's Disease Assessment Scale – Cognitive Subscale" (ADAS-Cog): testing verbal memory (immediately and delayed),
- "Stroop Color and Word incongruent score" (stroop): testing for interference in the reaction time when performing a task,
- "Mini Mental State Examination (MMSE)": global cognitive function, inclusion/exclusion criteria.

The intervention group proceeded to perform sixty-minute exercise sessions three times weekly for sixteen weeks while the control received usual care. The sixty minute program was composed of aerobic exercise with a number of

[4] Copenhagen, Slagelse, Roskilde, Odense, Aalborg, Aarhus, Svendborg and Glostrup.

Figure 3.3: A number of tools available for exercise during the aerobic exercise sessions each week.

strength building exercises. Furthermore, several exercises assessing both physical and functional performance were performed at baseline and at a 16-week follow-up (non-exhaustive list):

- 6-min Astrand Cycle Ergometer test: assessing maximal oxygen intake and average heartrate,
- Timed Up and Go (TUG),
- chair stand,
- Timed 10 minute walk, and timed 400 meter walk test,
- Dual task performance measured by performing the 10 minute walk test and respectively naming the months and counting from 50, backwards.

Figure 3.3 illustrates some of the exercise equipment that the participants had available for aerobic training during their exercise sessions each week.

At a sixteen week follow-up, both groups received cognitive testing again. The control group then proceeded to perform the exercise program for four weeks, whereas the intervention group received treatment as usual. Finally, both groups received a number of tests at a second follow-up at twenty weeks. The full flow starting from recruitment and screening and ending at the second follow-up has been illustrated in Figure 3.4.

The trial went on for five consecutive rounds, running from January 2012 until June 2014. During each round, all participants from Rigshospitalet, Roskilde and Odense[5] were invited to receive MR scans at Hvidovre Hospital twice, once at baseline, and once again at follow-up at sixteen weeks. The full details of

[5]Odense only participated in the MR study during round four and five.

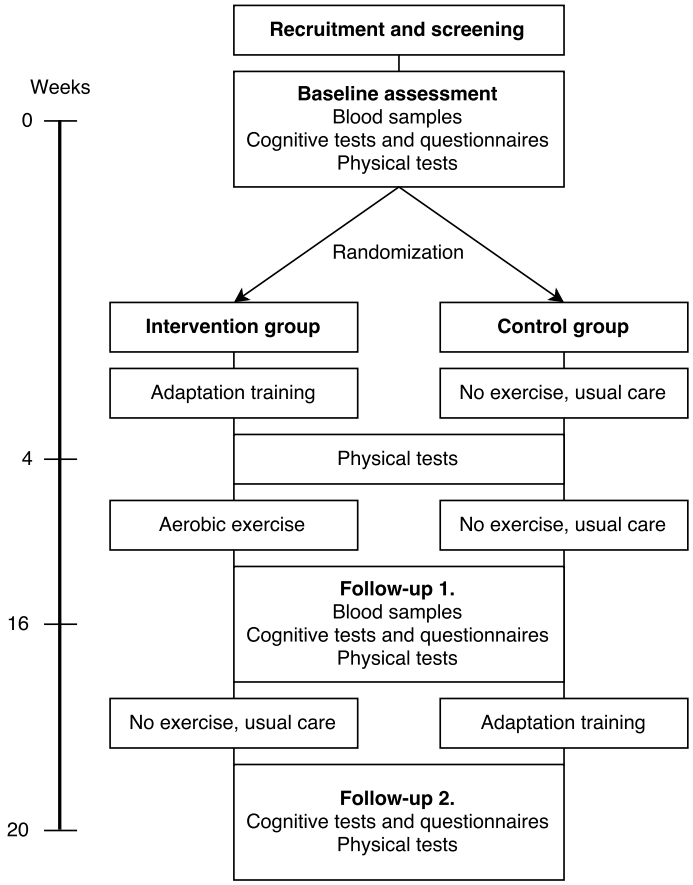


Figure 3.4: The ADEX study flow for a single round, starting with recruitment and screening and ending at the second follow-up.

the exercise study and its design were previously described in [Hoffmann et al. 2013].

Whereas the main ADEX study focused on the effects of exercise on cognition and physical performance, the MRI substudy focused on measuring brain changes in a subpopulation from the main study. For this purpose, Freesurfer [Fischl et al. 2002; Fischl 2012; Reuter et al. 2012] was employed, as it is an excellent tool to obtain measures of both cortical thickness and volume, as well as subcortical volume of e.g., the hippocampal subfields [Van Leemput et al. 2009].

3.3 Freesurfer - A Tool for Brain Morphometry

The term *morphometry* refers to "measurement of shape". Brain morphometry is correspondingly concerned with measuring volume and/or structural changes in the brain. The Freesurfer (FS) processing pipeline measures shape by segmenting the brain regions of interest (ROI), such as WM, GM and CSF, as well as underlying subcortical structures such as the hippocampus. Furthermore, WM and pial surfaces are also computed, and the cortex is finally parcellated[6] into regions, e.g., the sulci or gyri of the frontal cortex using some predefined regional, cortical atlas [Desikan et al. 2006; Destrieux et al. 2010].

The volume of these are then measured, i.e., by counting voxels, and potentially weighing these given probabilities of belonging to different structures, i.e., due to partial volume effects. One example of this is FS segmentation of the hippocampal subfields, as the hippocampus in particular is severely affected by partial volume effects. Using the obtained delineations of the WM and pial surfaces, cortical thickness can also be computed. As such, Freesurfer yields a wide range of results that are very easily interpretable from a morphometrical point of view. These can be used to do interesting statistics, such as comparing if the average size of the hippocampus differs between two groups.

3.3.1 The Cross-sectional Pipeline

Cross-sectional (single time point analysis) FS can be broken down into three stages, each of which are again composed by a number of steps. Full FS processing of a single time point is referred to as the *cross*. An overview of the stages and the most important steps in each will be presented in the following[7]. The pipeline has been illustrated by Figure 3.5.

Several of the FS steps work on an assumption of WM having a mean intensity around 110 (GM is generally assumed to be around 75). Given that FS needs to facilitate segmentation of a vast range of different T1-weighed images due to i.e., scanners made by different manufacturers (e.g., Siemens, Philips, General Electrics), of different models and field strengths (to date, FS supports $1.5T$ and $3T$), and differences in T1-weighed sequencing, FS performs several intensity normalization steps throughout.

[6] which is a different way of expressing segmentation.

[7] Based on the current stable Freesurfer version 5.3, which is fully described at `https://surfer.nmr.mgh.harvard.edu/fswiki/ReconAllTableStableV5.3`.

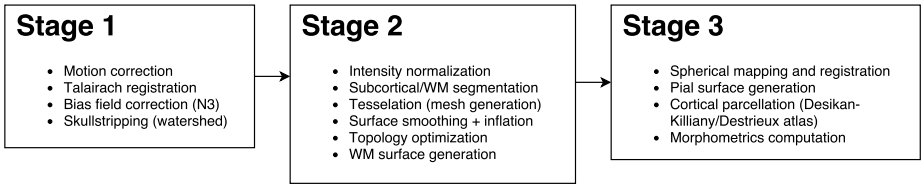


Figure 3.5: The Freesurfer 5.3 pipeline and its three stages, including the underlying major steps.

Stage 1 - Preparing the Data

This stage is concerned with preparing the data for segmentation. It involves performing motion correction (if several scans are available of the same subject) and conforming the data to $256 \times 256 \times 256$ isotropic 1 mm^3 voxels. The data is then bias field corrected using the N3 algorithm and skull-stripped using the watershed algorithm (which unless disabled, involves performing a registration to a brain template with skull in Talairach space to guide the skullstripping)[8].

Stage 2 - Computing Segmentations and Surfaces

In stage 2, the first step computes a transformation to a template space containing a probabilistic atlas over subcortical structures. The intensity of WM is then normalized using a number of predetermined WM control points in the atlas space. Following this, the initial registration is used to initialize a second one using the normalized data as input. A number of preparation steps are then run (e.g., removal of neck), and the data is finally segmented into subcortical structures using a Bayesian approach (generative model). The data is then normalized a second time using the previous normalization as input[9], and finally a segmentation of WM is computed[10].

After subcortical and WM segmentation, a filled WM volume is computed, which is used to initialize tessellation (generating a triangular mesh) of the WM segmentation. The mesh is then smoothed and inflated, given that the cortex is assumed to behave like a piece of paper which has been curled together, in order to resemble a brain. These assumptions are key in simplifying the computational analysis of surfaces and associated steps.

[8] Brainmask editing takes place here.
[9] Manual control point insertion takes place prior to this normalization
[10] Manual WM editing takes place after this step.

Topological errors (e.g., holes or triangles that cross each other) are then fixed using a number of steps and the final, corrected mesh is used to compute the final white matter surface, which is again smoothed and inflated in preparation for stage 3.

Stage 3 - Parcellating the Pial Surface and Computing Morphometrics

The already inflated surface is further inflated to resemble a sphere. This step is a prerequisite for the following registration to a cortical atlas space. Cortical regions following the Desikan-Killiany and Destrieux atlases are then mapped onto the surface. Finally, the pial surface itself is generated, and a broad range of morphomometrics are calculated, such as volume for subcortical structures and cortical, regional volume and thickness.

3.3.2 Longitudinal Analysis

FS also supports longitudinal analysis of several time points of the same subject. This is essentially an extension of the cross-sectional pipeline, where cross-sectionally processed time points are used to generate a common subject specific template, also commonly referred to as the *base*. All time point images are iteratively co-registered via rigid transformations to a voxel-wise median image, creating an unbiased subject reference space and template that represents the average subject anatomy across time [Reuter et al. 2010; Reuter et al. 2012]

The subject template image is then processed mainly with the regular FS pipeline, to localize and estimate average subject anatomy, which is then used for a common initialization of the subsequent processing (fine-tuning) of the individual time points [Reuter et al. 2012]. This common initialization greatly removes variability in measurements, such as the white matter segmentations or cortical surface construction, and significantly increases measurement reliability. Each of these final runs in each time points are referred to as the *long*.

The longitudinal approach has the advantage that bias due to asymmetry in the registration is avoided, as all timepoint analyses are performed within a common space of reference (the template), rather than by registering everything to the baseline. Consequences of processing and interpolation bias were previously discussed in e.g., [Fox et al. 2011; Thompson and Holland 2011]. An additional advantage of using a subject specific template for initialization is that manual editing (where necessary) can often be performed on the template image, rather in all or a selection of individual time points.

3.3.3 Editing

A significant part of the work done throughout this PhD study revolved around collaborating with as well as supervising two experienced readers at DRCMR for the purpose of using FS (editing) and Freeview (the FS inspection and editing software interface). This in turn required first formalizing a training environment (composed by a few of the initial ADEX scans), where both readers practiced FS editing before actual work began on the whole dataset. Throughout the ADEX project, multiple discussions and supervising sessions where held in order to help guide both readers in how to handle FS editing given particular problematic segmentations of some ADEX scans, which proved both difficult and time consuming due to the quality of the data.

This section describes the three manual editing operations that may be necessary in order to obtain proper subcortical segmentations as well as delineations of the pial and WM surfaces. All resources, including more detailed descriptions of the editing as well as the included illustrations, are all available online on the FS wiki[11].

Brainmask Editing

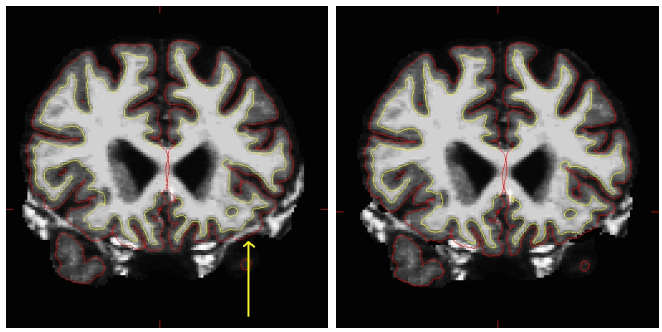
In cases of dura or skull that borders brain tissue without separating voxels (i.e., containing CSF), FS may fail to segment the pial surface correctly. This results in the pial surface being grown into the dura or skull. This error is corrected by editing the brainmask, which is effectively how FS limits how far the pial surface can be grown.

After editing the brainmask using an overlay of pial and WM surfaces as a guideline, and removing voxels that were incorrectly classified as cortex, FS can be rerun. At this point, the pial surface will be properly aligned. Often, it is not necessary to remove all of the 'offending' voxels, and FS delineates the surfaces correctly after just a few voxels have been removed. The process has been illustrated in Figure 3.6.

Control Point Insertion

In some cases FS fails to segment the WM properly, and barring problematic data due to e.g., lesions, this is typically a result of non-optimal normalization.

[11] `https://surfer.nmr.mgh.harvard.edu/fswiki/FreeviewGuide/FreeviewWorkingWithData/FreeviewEditingaRecon`



(a) (b)

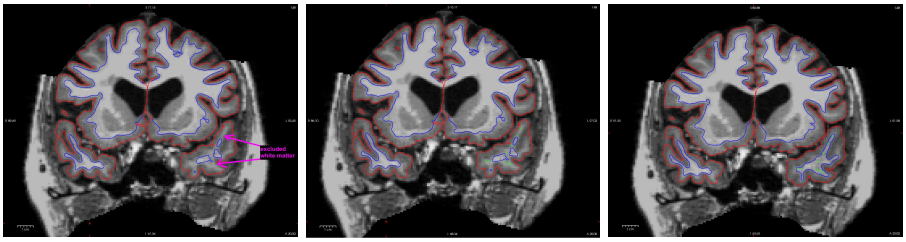
Figure 3.6: The pial surface (red outline) has not been delineated properly because bordering skull/dura leads to neighboring voxels with similar intensities, leading to error in segmentation of voxels. By editing the brainmask and rerunning FS from that stage, the pial surface aligns itself nicely. a) Before editing the brainmask, b) after FS processing.

The solution is to insert a number of control points (CP) in voxels of WM in the problematic region. The intensities in the voxels where CP's where inserted helps to steer the overall normalization, and often the segmentation will be more accurate after processing the data again.

There are cases where CP insertion will result in a failure to improve normalization and the following segmentations. One, if CP's are inserted in GM, these voxels will be considered WM and the contrast betwen the two tissue types will be washed away after normalization, effectively ruining the following segmentations. Two, if contrast is already bad, due to e.g., severe lesions, FS will already have a hard time to normalize the data, which means CP insertion will not help. The reason is that the problematic voxels *look like* GM voxels due to their low intensity, even though the human eye will safely classify them as WM. CP insertion has been illustrated in Figure 3.7.

White Matter Segmentation

Lesions can be particular problematic for FS, because they sometimes resemble GM in T1-weighted images. This often results in subcortical voxels being segmented as cortex, leading to WM and even pial surface delineations deep within WM. The problem has been illustrated in Figure 3.8. The solution is to *fill* these voxels in the WM segmentation, effectively making them appear as if they are WM. Of course, this is not correct biologically, but since FS does not



(a) (b) (c)

Figure 3.7: WM intensities have not been properly normalized, which results in a WM surface (blue outline) that is inaccurately delineated. One typical cause of this is when the bias field has not been corrected properly, or if the data has poor contrast. Control points are inserted, FS is rerun from that point in the pipeline, and the problem is solved. a) before insertion of control points, b) after insertion but before FS processing, c) after FS processing.

segment lesions[12], treating such voxels as WM is the only way for the program to generate the WM and pial surfaces successfully.

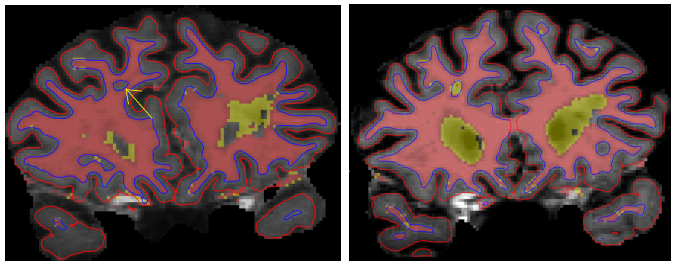
A problem with a similar outcome is when skull is incorrectly segmented as WM outside of the brain. The problem has been illustrated in Figure 3.9. As before, the solution is to remove the offending voxels in the WM segmentation, and then rerun FS. This problem can also be handled by editing the brainmask as previously described. However, since the problem here is principally different from that in the brainmask case, it is technically more correct to edit the WM. It is not clear if handling the problem in one way or the other has any noteworthy effects on the final segmentations.

Longitudinal Editing

Editing in longitudinal FS processing is essentially the same as is done for single time points, with the advantage that it is often sufficient to simply edit the subject-template (*base*) which affects all time points in the subsequent long runs. A detailed overview of how to do these edits, and what can be skipped, is available on the FS wiki online[13]. After edits to the subject-template (*base*) and in some situations the initial *cross* runs, no further editing should be necessary in the final *long* runs.

[12]FS does label dark voxels within subcortical structures as *hypointensities*. While this is a trademark of lesions in T1-weighed images, it *should not* be confused with lesion segmentation.

[13]`https://surfer.nmr.mgh.harvard.edu/fswiki/LongitudinalEdits`



(a) (b)

Figure 3.8: a) A lesion in the brain leads to voxels which are (correctly) identified as non-WM. However, the hole is segmented as GM, which leads to incorrect delineation of the WM (blue outline) and pial (red outline) surfaces. b) To correct this, the hole is filled out in the WM segmentation (semi-transparent red color).

In the ADEX project, we chose a more conservative approach of editing brainmasks and control points in the *crosses*, omitting the white matter segmentation as this is recomputed in the *base*. Then, further editing of brainmasks, control points and white matter segmentation took place in the *base* if necessary. This is generally recommended, as high quality segmentation in the *cross* results in better initialization of the *base*, which in general should translate to fewer edits necessary in the *base*.

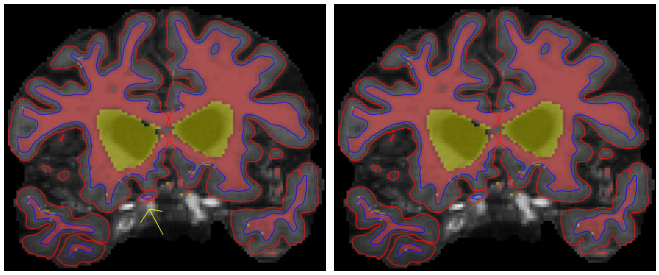
3.4 Contributions

This section discusses the main contributions done in the ADEX MRI study.

3.4.1 ADEX MRI Study Outcomes

The main hypothesis of the ADEX MRI study was that moderate-to-high aerobic exercise would have beneficiary effects in areas of the brain known to be affected by AD in a population of patients with mild-to-moderate AD. In particular in the hippocampus, where exercise has previously been shown to be effective in a population of healthy elderly [Erickson et al. 2011], but also on cortical regional thickness.

As presented in paper A, it was not possible to show that the changes in brain



(a) (b)

Figure 3.9: An example of FS failing to delineate the pial surface properly because bordering skull/dura leads to neighboring voxels with similar intensities. The separation is consequently not detected by FS properly. By editing the brainmask and rerunning FS from that stage, the pial surface aligns itself nicely. a) Before editing the WM, b) after editing the WM, but before generating new surfaces.

measures (subcortical volume, cortical thickness) were significantly different between the control and intervention groups over a 16 week period. One possible explanation for this is the duration of the study; 16 weeks may be too little to differentiate groups based on atrophy rates. This seems likely when considering that a) [Erickson et al. 2011] showed effects, but after a full 1-2 years and b) the effects where shown in a healthy population. Another possibility is that the study suffers from bad data quality and too few participants in the MRI subpopulation, as 13 subjects left the study prematurely, 9 were excluded due to poor MRI data quality (severe motion artifacts), 6 were further excluded due to problems processing the data with FS and finally 1 was excluded due to error in processing outcomes. Even so, there's a number of points that arguably qualifies the study as a success.

First, it is the first study (to our knowledge) of its kind to explore the effects of exercise in a population of patients with AD using MRI techniques. The significance of this is the establishment of a reference for future exploratory studies on the effects of exercise in AD (possibly also other pathologies). Given that the exercise duration is a likely reason for the study outcomes, it would be highly interesting to pursue a similar study where the effects of exercise were recorded over a longer period, e.g., 1-2 years.

Second, the MRI study did show significant, positive correlations between how efficiently the exercise group did their aerobic exercise[14] and both changes in hippocampal volume and frontal cortical thickness. This finding supports pre-

[14]See the paper for details

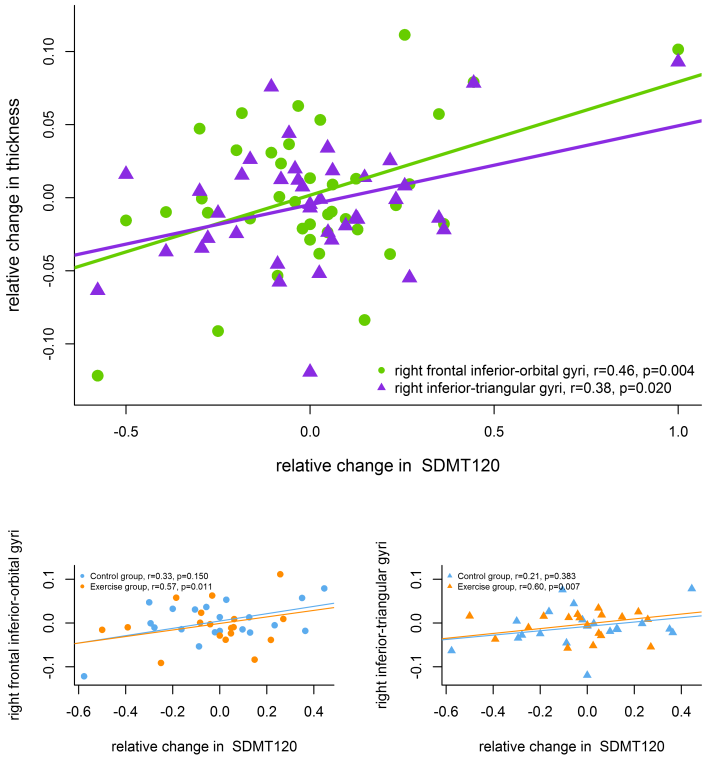


Figure 3.10: Largest positive correlations between changes in the frontal cortex and the SDMT cognitive test scores.

vious studies on the effects of exercise on brain growth [Colcombe et al. 2003; Colcombe et al. 2006; Ruscheweyh et al. 2011] in healthy elderly, and easily complements the argument of 16 weeks being too little: "exercise does work, you just need to do it longer".

Third, separate from the effects of exercise, significant, positive correlations were shown between changes in frontal, cortical sulci and gyri thickness and changes in cognitive performance scores for mental attention (SDMT) and verbal fluency (VFT). The frontal cortex is known to associate with these types of cognition, and as such, the finding agrees with previous literature. The two largest correlations for specific gyri and sulci of the frontal cortical gyri and sulci thicknesses, and the cognitive SDMT and VFT test scores respectively have been illustrated in Figures 3.10 and 3.11.

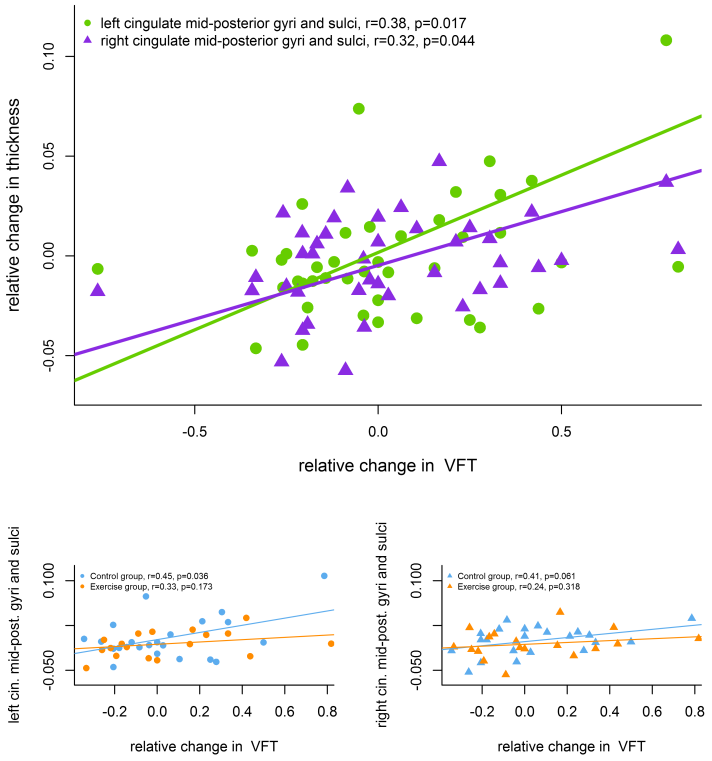
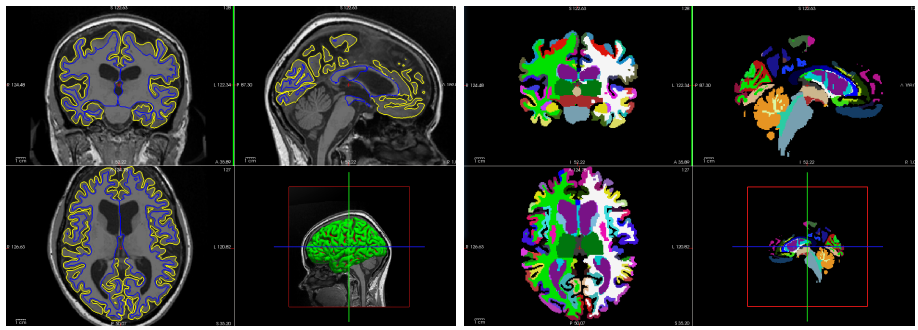


Figure 3.11: Largest positive correlations between changes in the frontal cortex and the VFT cognitive test score.

Somewhat pragmatically, it can be argued by inference given two and three that since exercise does seem to have an effect on brain volume (including frontal, cortical thickness), and since changes in this area positively correlates with cognition, stimulating brain growth by means of exercise *should* have a positive effect on cognition. Even if this relationship hasn't been proven directly in this study, the message is clear: "go out there, and get that exercise done".

3.4.2 Formalizing the use of Freesurfer at DRCMR

This study was the first to use a full, extensive, longitudinal FS analysis at DRCMR. This has resulted in a lot of experience using the tool, which may help



(a) (b)

Figure 3.12: a) Pial (yellow outline) and WM (blue outline) surfaces displayed with coronal (top left), axial (bottom left) and sagittal (top right) views, together with a 3D representation of the pial surface (bottom right). b) The same brain displayed with labels in different colors for subcortical structures and cortical regions (Destrieux atlas). The brain is visibly affected by atrophy in the cortex with visibly enlarged ventricles.

to encourage its use in future studies and to improve the quality and reliability of reported measures. Figure 3.12 shows segmentations and surfaces obtained using FS for one patient in the ADEX study.

More practically, part of the contribution also involved teaching and supervising relevant staff in its use, in particular collaborating with and teaching two clinicians that did all editing during the FS data processing.

3.5 Discussion

This section presents some considerations about the employed methodology in the ADEX study, in particular emphasizing the importance of having a proper "toolbox", and perhaps even more important, using the tools correctly. Throughout the section, interesting topics for future work are suggested.

3.5.1 Alternative Techniques for Brain Morphometry

Voxel-based morphometry (VBM) can (to some extent) be considered an alternative to FS-based analysis, and is done by statistically comparing the image

volume at each voxel within a region of interest (ROI) following a registration into some common template space, rather than measuring e.g., volume of a whole structure such as the hippocampus. Which method to use depends on preference, but more impotantly on what the goal of the study is.

VBM as a technique is highly dependent on the methods chosen for registration, and (as previously mentioned) if these are not accurate, and in particular if they do not properly account for asymmetry bias, findings may be significantly affected [Bookstein 2001]. However, considering that these problems have been addressed in recent work, this is likely less of an issue today. Speaking in favor of VBM is the fact that good registration can be obtained relatively fast today. This is advantageous in studies with large populations, potentially with multiple time points available per subject.

One of the merits of Freesurfer is its ability to get good measures of both structural volume and cortical thickness. FS has shown to have at least comparable accuracy to that of "ground truth" manual segmentations and delineations of brain structures and surfaces. Furthermore, given that the pipeline is automated (barring the manual edits that can be necessary when data quality is bad), it does not suffer from the variation that is introduced when different people annotate the brain image. The downside is that it takes FS between 12 and 16 hours to process a single volume on a modern PC.

Here, FS was chosen because it aligned really well with "both sides" of the PhD study, as e.g., both hippocampal (subfield) volume and cortical thickness was of interest, and because bias field correction and its application in FS were to be studied. Future VBM analysis of the ADEX data could be of interest, but given the inability to show differences between groups with FS, and also given the size of the population as well as the data quality (which increases the risk of error in registration to the template space), it is doubtful any noteworthy effects would be discovered.

Another reason to use FS for data analysis is that it generates actual surfaces which can be used for surface-based analysis, something which VBM does not do. Further considering that the longitudinal analysis presents an abundance of accurate segmentations within the same space of reference for the time points of each subject, FS lends itself well to extending the study into areas which depend on these ROI to be defined. As an example, a study of the ADEX data measuring tissue perfusion (blood flow) using arterial spin Labeling (ASL) has begun at DRCMR. This study can utilize the readily available FS segmentations at no additional expense.

3.5.2 Freesurfer: a Supervised, Generative Segmentation Method

Interestingly, FS segmentation of subcortical structures (e.g., WM) is based on a generative modeling approach (such models in general are elaborated upon in chapter 5), where some of the associated parameters, i.e., mean values of WM and GM, have fixed values. Consequently, for the pipeline to work successfully the tissues must have their intensities normalized to these values (previously described), which in turn depends on accurate placement of CPs (both manually and from the atlas).

The reason why this is interesting is because it is not difficult to estimate all parameters, including mean tissue intensities, when a generative model is employed to segment the data, in particular when a probabilistic atlas is already available for the tissues or subcortical structures. In FS, such an atlas as well as the necessary transformation between the image and the atlas space *is* available, and to some extent already used. Works were all model parameters are estimated from the data are e.g., [Van Leemput et al. 1999b; Zhang et al. 2001].

Furthermore, the work by [Ashburner and Friston 2005] showed how atlas registration, bias field correction and tissue classification can be combined into one unified model. As will be clear in the following chapter, this model is intimately related to the ones employed for the purpose of bias field correction in this thesis. For now, it is sufficient to mention that FS might very well benefit from transitioning from decoupled registration, bias field correction and segmentation steps, into a more unified approach. This is currently being considered for future releases of the pipeline.

3.5.3 The Importance of Proper Bias Field Correction

The N3 algorithm depends on a distance hyper-parameter, which defines how smooth the bias field estimate should be. In FS, this hyper-parameter is controlled using a "-3T" hyper-parameter, which should always be enabled when processing 3T data [Boyes et al. 2008]. In the process of running longitudinal FS, it came to our attention that FS does not inherit the 3T N3 hyper-parameters from the cross-sectional analysis. Before we discovered this, the base template generation and longitudinal analysis were consequently run using hyper-parameters tuned for 1.5T.

Whereas results using the 1.5T bias field correction hyper-parameters showed no significant effects or correlations in the ADEX study, the picture changed

drastically when the 3T hyper-parameters where introduced and FS rerun, which lead to the results discussed in section 3.4.1 as well as the included paper. This serves to highlight just how important good bias field correction is, and consequently motivates why research into better models for bias field correction methods should be pursued. While the work performed on bias field correction models (unfortunately) did not reach a stage where it could be included in clinicial studies (such as the ADEX study), this research is the focus of chapter 5.

Chapter 4

Bias Field Correction Literature and Validation

In chapter 5, we will mainly discuss a particular generative modeling approach to bias field correction. We also touch briefly upon the N3 algorithm, which is the *de facto* standard for bias field correction and has the advantage that no a-priori information about the image is necessary. However, many other methods for bias field correction exist. Overall, we can partition these methods into three distinct model groups, although closer inspection reveals that they are, in fact, quite similar:

- **Generative** model-based methods seek to maximize the posterior probability of a set of unobserved variables (including the bias field) given the image. These models will be covered more in depth in chapter 5.
- **Heuristic** methods seek to estimate the bias field using more "ad hoc" approaches such as image filtering, surface fitting or histogram matching. These methods typically make strong assumptions about how the bias field affects the image.
- **Hybrid** methods fall somewhere in between; they can be shown to employ or relate to an underlying probabilistic model, but may not optimally optimize the involved parameters.

In the following sections 4.1, 4.2 and 4.3 we discuss some of the more well-known and/or recent methods. We then progress to discuss a number of longitudinal models for bias field correction when several time point scans of the same subject are available in section 4.4. In the final Section 4.5, we discuss a number of popular ways to validate bias field correction, also covered within the presented literature. The overview presented here is not exhaustive. For a more complete or supplementary overview, a number of works review the literature [Velthuizen et al. 1998; Styner and Van Leemput 2004; Belaroussi et al. 2006; Vovk et al. 2007] and/or evaluate performance for a number of bias field correction methods [Arnold et al. 2001].

4.1 Generative Model-based methods

As previously described, methods that use these models are probabilistic and explicitly try to maximize an objective function which describes the posterior probability of observing the data. This typically involves representing the underlying uncorrupted data with a mixture of Gaussians model, where each voxel is assumed to originate from some unknown label (e.g., a tissue class). The bias field is represented by a linear combination of basis functions, which in the case of EM-based optimization is fit to the residual data, obtained by subtracting the uncorrupted data estimate from the observed data. This specific approach will be discussed in depth in chapter 5.

[W. M. Wells et al. 1996] combines a model for tissue classification and bias field correction where means and variances for a number of tissue classes are assumed known. The data is log-transformed prior to parameter estimation such that the bias field can be assumed additive, and optimization is performed using an expectation maximization (EM) algorithm.

Given the assumptions of a-priori known label means and variances, a training phase is necessary prior to correction. [Held et al. 1997] further implement a Markov random field (MRF) model on the labels in order to minimize the effect of noise in the tissue classification. This (and the use of MRF models in general, within a generative framework) results in a complication of the EM parameter optimization. Similarly, [Guillemaud and Brady 1997] revise the model by [W. M. Wells et al. 1996] such that voxel intensities that are not likely to belong to any of the tissue labels (represented by a "garbage" label) are modeled by a uniform distribution, scaled by a user defined parameter. This results in bias field estimation that only takes place with respect to voxels that fall within tissue labels. However, as before the method suffers from a dependency on a training phase.

[Van Leemput et al. 1999a] show how bias field correction and tissue classification can be integrated in a three-step algorithm where all parameters of the model are fully estimated. They further show how to extend the model to include multi-channel data, and how to integrate filtering of voxels containing only background noise, which are typically removed prior to correction using data preprocessing. The noise is assumed to follow a Rayleigh distribution [Gudbjartsson and Patz 1995]. Finally, they show how to account for slice-by-slice constant offsets in multi-sclice 2D MRI images. As before, the data is log-transformed such that the bias field can be modeled as an additive effect, here using a linear combination of polynomials. Optimization is performed using a generalized expectation maximization algorithm (GEM) due to interdependent model parameters.

[Ashburner and Friston 2005] show how image registration using a deformable tissue atlas can be combined with tissue classification and bias field correction in one, unified model. Here, parameter estimation is performed in the original data domain using a combination of EM (mixture model) and the Levenberg-Marquardt algorithm (registration, bias field correction). The bias field is here modeled using a linear combination of cosine basis functions.

4.2 Heuristic Methods

A number of methods attempts to remove low-frequency components in the image, assumed to be the bias field effect, by means of low-pass filtering techniques. Some of the published literature on such methods are [Axel et al. 1987; Brinkmann et al. 1998; Cohen et al. 2000].

Some methods seek to estimate the field by fitting spatially smooth basis functions to the image data directly, i.e., thin plate splines to a number of image reference points selected by the user within a tissue [Dawant et al. 1993], or fourth-order polynomials to a collection of homogeneous regions obtained using an iterative thresholding of image gradients, which are then combined to compute and obtain a global fit for all regions [Meyer et al. 1995]. [Brechbühler et al. 1996] obtains a bias field estimate using a second-order Legendre polynomial representation of the bias field. The estimate is found by minimizing an energy function over the residual of the observed data minus the bias field estimate and a number of class means, which are predetermined by the user.

[Likar et al. 2001] seeks to minimize the entropy of the bias field corrupted image taking into account both multiplicative bias and additive noise, describing both using linear combinations of normalized polynomials respectively. The method

relies on a mean intensity-preserving condition between the uncorrected and corrected image, and optimal parameters are those that minimize the entropy, using a combination of optimization techniques (Powell's and Brent's methods). Similarly, [Mangin 2000] follows a similar approach, but only consider the multiplicative field which is modeled using splines with adaptable control points. Optimization is performed using a stochastic algorithm that relies on a fast annealing schedule.

[Li et al. 2009] presents a variational level set approach to bias field correction and segmentation, which utilizes a k-means clustering algorithm to partition the data into regions in the image domain. The method works under the assumption that the regions are separable, such that the bias field can be estimated independently within each region. Following estimation, the estimates are combined in order to obtain a global bias field.

More recently, [Adhikari et al. 2014] seek to estimate the bias field of 2D MRI images by fitting a Gaussian surface to each of the gradient maps for a number of homogeneous intensity regions, which are identified by identifying image histogram peaks. The bias field estimate is then obtained by taking the average of the Gaussian surfaces.

None of these approaches differentiate between smooth variations in the image due to biological variance within a tissue or the bias field, which makes them susceptible to removing biological information from the image.

4.3 Hybrid methods

The N3 algorithm [Sled et al. 1998] falls somewhere in between the generative model-based and heuristic methods. The method claims to be non-parametric, but as shown in [Larsen et al. 2014], the method does, in fact, employ a parametric, generative model where some of the involved parameters are estimated using a heuristic updating scheme. As previously discussed, the method has the advantage that it can be applied to any MRI without prior information about the image.

[Tustison et al. 2010] present N4ITK as an evolution of N3, which replaces the cubic b-spline smoothing scheme from N3 with a more elaborate scheme where control points are allowed to adapt to the image. Interestingly, from a generative model point of view, the updates for the bias field coefficients are the only part of N3 that is "correct", relative to the model. Given that N4ITK replaces the smoothing scheme, this model should be considered closer to the

heuristic models than N3. Whether the more heuristic approach of N4ITK is an advantage or not, is yet to be fully explored.

[Pham and Prince 1999] presents a fuzzy segmentation scheme that combines tissue classification with bias field correction. The method seeks to minimize an objective function defined as the two-norm between voxel and class intensities weighed with a membership value. Interestingly, this approach is very similar to generative models, as the membership values bears resemblance to the posterior probabilities of class assignments in generative models.

Similarly, [Ahmed et al. 2002; Liew and Yan 2003; Ji et al. 2011] modify or extend the fuzzy c-means segmentation scheme in order to improve performance, but otherwise preserve the core scheme of the method. Some of these methods appear quite heuristic in their respective paper presentations, but are nevertheless listed here for consistency with respect to the fuzzy c-means algorithm. Note that the work by [Li et al. 2009] previously described bears a lot of resemblance to these methods. However, it seems that they try to distance themselves from the fuzzy c-means algorithm, for which reason they have been listed separately.

[Shattuck et al. 2001] formulates a bias field estimator using the conditional probability of observing the data given the bias field effect and a number of global tissue parameters (mean and variance) which are estimated by automated analysis of the image histogram. Based on an assumption of a (small) regionally constant bias field, sample bias field values are then obtained within uniformly positioned regions over the image, by minimizing a cost function on the residual between the observed regional data and corresponding "true" data histograms. These sample values are then smoothed to obtain a global bias field estimate using a regularized least squares fit of cubic b-splines that have been penalized on their bending energy.

4.4 Longitudinal Model-based Methods

The previous methods only consider models for cross-sectional bias field correction – cases where only one time point scan of the subject is available. More recently, attention has shifted towards how the bias field artifact is best estimated for longitudinal data, i.e., when there is more than one time point scan of the same subject.

[Lewis and Fox 2004] presents a model that corrects for the differential bias field between time point scans for the same subject. The method assumes a pairwise rigid registration before difference bias field estimation. The method does not

account for the bias field which is common to the subject time point scans.

Similarly, [Modat et al. 2010] introduces a model for correction of the difference bias field between time point scans within a non-rigid registration framework, using normalized mutual information as the optimization metric.

[Ashburner and Ridgway 2013] present a solution where diffeomorphic and rigid-body registration to an anatomical tissue atlas is combined with differential bias field estimation, following the spirit of a unified model for parameter estimationas as presented in [Ashburner and Friston 2005].

4.5 Validation of Bias Field Correction Performance

Validation of any bias field correction method is generally challenging, as no ground truth bias field is available unless synthetic or simulated images are used. One popular extensible hybrid Bloch equation and tissue template simulation based MRI simulator was presented by [Collins et al. 1998; Kwan et al. 1999]. This simulator is used to synthesize the images that are available in the online database "Brainweb"[1]. These images were used to train and validate the default N3 algorithm parameters and performance.

Most of these MRI simulators typically depend on precomputed bias fields which are then scaled up or down to simulate differences in field strength. Alternatively, the simulators make assumptions about or approximations to the MR physics. In the event of synthetically generated data, the bias field effect is typically modeled as a simple multiplicative effect given a sum of only a few spatially smooth basis functions, e.g., cosines, which is then applied to a ground truth template image. In most cases, these simulated bias fields are at best crude approximations to the bias field effect in real MRI, which makes their use in bias field correction performance validation questionable. It should be noted that at least one simulator [Stöcker et al. 2010] is known to produce very accurate MRI simulations. However, the computations are extensive, and to our knowledge requires several hours of computation in order to produce just one simulated 3D MRI image.

In any case, considering that MR images typically receive bias field correction as a preprocessing step in an automated segmentation pipeline[2], it is more

[1] `http://brainweb.bic.mni.mcgill.ca/brainweb/`.

[2] Unless it is jointly estimated during segmentation in a unified model.

relevant to look at quantitative performance metrics that can be used on real images to determine correction quality indirectly, or alternatively by looking at segmentation performance explicitly.

4.5.1 Indirect Measures

Relevant quantitative metrics are the coefficient of variation (CV) which is typically measured within WM in brain MRI, as this tissue is assumed to be very homogeneous. The CV is defined as the standard deviation over the mean

$$CV = \frac{\sigma}{\mu}. \tag{4.1}$$

Whereas the CV in WM may be a good measure of homogeneity, it does not consider how well the intensities in WM and GM separate, which is of much greater importance for automated segmentation algorithms. The coefficient of joint variation (CJV) [Likar et al. 2001] takes this into account by relating the standard deviation of two sets of intensities to their respective means

$$CJV = \frac{\sigma_1 + \sigma_2}{|\mu_1 - \mu_2|}, \tag{4.2}$$

e.g., WM and GM in brain MRI. The CJV between WM and GM is the primary quantitative performance measure for bias field correction used in this thesis. Other indirect performance metrics include the entropy E of the normalized image histogram H

$$E = -\sum_{n=1}^{N} H_n \log(H_n). \tag{4.3}$$

Any bias field correction algorithm should lower the entropy of the image, as the bias field increases the variance of the image intensities within homogeneous tissue (which leads to increased entropy). However, entropy is non-trivial to use as a performance metric when comparing methods, as it relies heavily on the binning of the histogram; slight variations in the distributions of the intensities (peaks and span) may affect the histogram such that the binning does not capture the shape of the distribution properly. Finally, just as the CV measure, entropy does not consider separation between tissue intensities.

4.5.2 Segmentation Performance

The Dice metric is a popular way to evaluate segmentation performance, which can also be used to measure bias field correction quality. The data is preprocessed with a bias field correction method, and automated segmentations are

then obtained. These are then compared to some ground truth segmentation, typically obtained manually by human experts. The Dice metric is defined as

$$S = 2\frac{n(A \cap B)}{n(A) + n(B)}, \tag{4.4}$$

where $n(x)$ denotes the number of voxels belonging to structure x, and A and B are the automated and manual segmentations respectively.

Cortical thickness can be used as a measure of robustness if several scans are available of the same subject, given the assumption that a good bias field correction algorithm should produce similar images of the same subject, after correction. In this thesis, we use Freesurfer [Fischl 2012] to obtain both measures of cortical thickness as well as segmentations of subcortical structures and tissue.

Chapter 5

Generative Bias Field Correction Models

This chapter introduces the reader to the underlying theory that is necessary to understand the relationship between generative (Bayesian) modeling and the popular N3 bias field correction algorithm, as presented in paper B, and further the generative framework for bias field correction that is presented in paper C. Section 5.1 introduces the reader to fundamental concepts in generative modeling. This involves providing a number of examples that shows how data can be generated from said models, and these are then extended to account for data that has been affected by a bias field. Section 5.2 describes how parameters of the generative models can be estimated, which is based on the expectation-maximization (EM) algorithm. In section 5.3 presents a formalized approach for designing an algorithm for bias field correction, and extends parameter estimation to the cases where generalized expectation-maximization (GEM) is necessary. Section 5.4 extends the presented generative modeling framework to include cases where several time points of the same subject is available. Section 5.5 presents the contributions made within generative modeling of bias field correction in this study, including papers B and C, and Section 5.6 concludes the chapter by discussing a number of topics related to these two papers, as well as potential for future work.

5.1 Generative Modeling

This section first establishes the basic concept of a probability density function. A number of generative models are then explained, exemplified and subsequently extended, starting with a simple example using a single Gaussian distribution, that is then extended to a Gaussian mixture which can describe an MR image. This model is then further extended to include the bias field artifact. Finally, this leads to a description of the full generative framework for bias field correction which is presented in paper C.

5.1.1 Probability Density Functions

Consider some continuous, random variable d. The probability for d to take a certain value can be expressed in terms of a probability density function $p(d|\boldsymbol{\theta})$ (PDF), where $\boldsymbol{\theta}$ defines a number of parameters (possibly none) that shapes the distribution of d. In order for this function to be a proper PDF it must be normalized, such that its integral is equal to one

$$\int p(d|\boldsymbol{\theta})\,\mathrm{d}d = 1. \tag{5.1}$$

The definite integral of the PDF over the interval $d \in [a, b]$ defines the probability for d to fall within that interval

$$P(d \in [a, b]|\boldsymbol{\theta}) = \int_a^b p(d|\boldsymbol{\theta})\,\mathrm{d}d. \tag{5.2}$$

If the interval is defined only in terms of b such that $P(d \in [-\infty, b]|\boldsymbol{\theta})$, we refer to this as the cumulative distribution function (CDF).

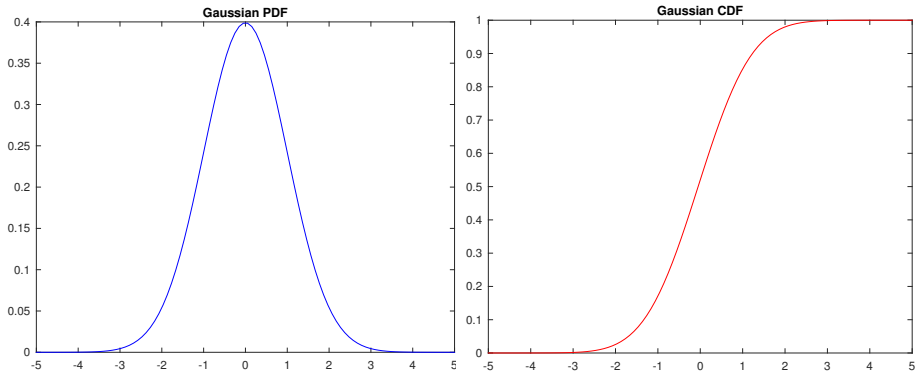
Here, we rely exclusively on the Gaussian (normal) distribution with PDF

$$\mathcal{N}(d|\boldsymbol{\theta}_d) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(d-\mu)^2}{2\sigma^2}\right), \tag{5.3}$$

which depends on two parameters, the mean and variance $\boldsymbol{\theta}_d = (\mu, \sigma^2)$ of the distribution. The PDF and CDF of a Gaussian distribution with predetermined parameters have been illustrated in Figure 5.1.

5.1.2 A Basic Generative Model

As implied by its name, a generative model randomly generates data d. This data which is *observable* depends on parameters $\boldsymbol{\theta}_d$ that are typically *unob-*



(a) (b)

Figure 5.1: A Gaussian probability density function with mean $\mu = 0$ and variance $\sigma^2 = 1$ (a), and its cumulative distribution function (b).

served (hidden). The generative model specifies the joint distribution over the observations and parameters, which for a single observation (sample) d is given by

$$p(d, \boldsymbol{\theta}_d) = p(d|\boldsymbol{\theta}_d)p(\boldsymbol{\theta}_d), \tag{5.4}$$

which is composed of the *likelihood* of observing data given the (hidden) parameters, and the *prior* probability of observing the parameters.

Given a vector $\boldsymbol{d} = (d_i, \ldots, d_N)$ containing a number of N samples from the model, and assume that they are conditionally independent given the parameters, the generative model is

$$p(\boldsymbol{d}, \boldsymbol{\theta}_d) = \prod_{i=1}^{N} p(d_i|\boldsymbol{\theta}_d)p(\boldsymbol{\theta}_d). \tag{5.5}$$

This implies that if the parameters $\boldsymbol{\theta}_d$ have been fully observed for some distribution of observable data, the shape of the histogram of the generated data will approach the observed data given enough samples.

As will be clear in the section 5.2, the generative framework is very powerful, as it allows us to estimate values for the unobserved parameters $\boldsymbol{\theta}_d$. For now, we assume that all parameters have been fully observed (denoted $\hat{\boldsymbol{\theta}}_d$) which corresponds to letting the distribution of $p(\boldsymbol{\theta}_d)$ be generated by a delta function

$$p(\boldsymbol{\theta}_d) = \delta(\boldsymbol{\theta}_d - \hat{\boldsymbol{\theta}}_d), \quad \delta(x) = \begin{cases} \infty, & x = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{5.6}$$

This simplifies following examples where we demonstrate how data is generated by means of sampling.

Histogram Approximation

For a number of ordered points $[z_1, z_2 \dots z_{M+1}]$ that defines M intervals, a histogram is the count of occurrences that is observed for the random variable d within each interval for a total of N observations (samples) of d

$$H_j = \sum_{d \in [z_j, z_{j+1}]} 1, j = 1, \dots, M. \tag{5.7}$$

The normalized histogram can be considered an approximation to the probability density function given the definite integral over M intervals spanned by the set of ordered points $z_1, z_2, \dots, z_{M+1}$

$$\tilde{H}_j = \frac{H_j}{\sum_{j'=1}^{M} H_{j'}} \approx P_j = \int_{z_j}^{z_{j+1}} p(d|\boldsymbol{\theta}) \, \mathrm{d}d, \tag{5.8}$$

assuming that the span of $[z_1, z_2 \dots z_{M+1}]$ is sufficiently wide, such that

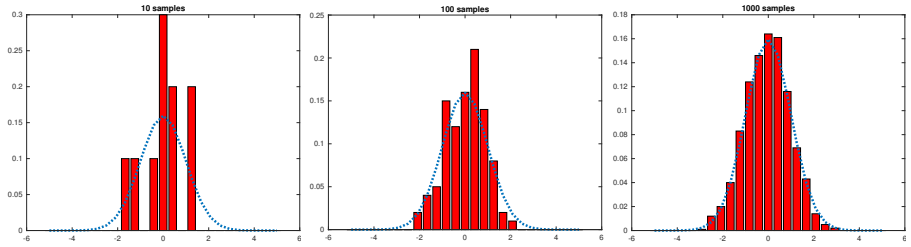
$$\sum_{j=1}^{M} P_j \approx 1. \tag{5.9}$$

If we draw enough samples, their distribution will be statistically similar to that of the PDF, as illustrated by the histogram. This has been illustrated for the same Gaussian distribution as in the previous example, using $N = 10$, $N = 100$ and $N = 1000$ samples in Figure 5.2. The actual value of d we generate is practically given by e.g., the center of the bin it falls within for each sample.

5.1.3 The Gaussian Mixture Model

We can model distributions of more complex data by using a linear superposition (mixture) of L Gaussians, each with its own μ_l and variance σ_l^2. In order to use this model, we first need to introduce the concept of labels.

We consider a scenario where each voxel $i \in \{1, \dots, N\}$ in the image $\boldsymbol{d} = (d_1, \dots, d_N)^T$ belongs to one of L possible labels $l_i \in \{1, \dots, L\}$ with some



(a) (b) (c)

Figure 5.2: a) The histogram of a Gaussian distribution obtained using 10 samples (a), 100 samples (b) and 1000 samples (c). As more samples are used, the shape of the histogram resembles the Gaussian distribution better.

certainty. I.e., the probability for observing a label image denoted $\boldsymbol{l}$ where all voxels belong to the same label, and given some parameters $\boldsymbol{\theta}_l$ is

$$p(\boldsymbol{l}|\boldsymbol{\theta}_l) = \prod_{i=1}^{N} p(l_i|\boldsymbol{\theta}_l), \tag{5.10}$$

where we have assumed that the occurrence of a label in a voxel is conditionally independent from the rest. We further consider the case where a given label occurs with the same relative frequency π_l in each voxel

$$p(l_i|\boldsymbol{\theta}_l) = \pi_{l_i}, \tag{5.11}$$

i.e., the probability for observing l_i is given by a discrete number of L probabilities stored in the parameter vector $\boldsymbol{\theta}_l = (\pi_1, \ldots, \pi_L)$, each of which satisfies $0 \leq \pi_l \leq 1$ and together $\sum_{l=1}^{L} \pi_l = 1$.

We assume that the intensity in a voxel is generated from a Gaussian distribution associated with label l_i

$$p(d|l_i, \boldsymbol{\theta}_d) = \mathcal{N}(d_i|\mu_{l_i}, \sigma_{l_i}^2), \tag{5.12}$$

which results in the following likelihood for observing the data $\boldsymbol{d}$ given an image composed of the same label $\boldsymbol{l}$, as well as the associated Gaussian parameters $\boldsymbol{\theta}_d$

$$p(\boldsymbol{d}|\boldsymbol{l}, \boldsymbol{\theta}_d) = \prod_{i=1}^{N} p(d_i|l_i, \boldsymbol{\theta}_d) \tag{5.13}$$

where again we have assumed conditional independence between voxels, i.e., given some label l_i the model generates intensities in a voxel independently of the labels in other voxels.

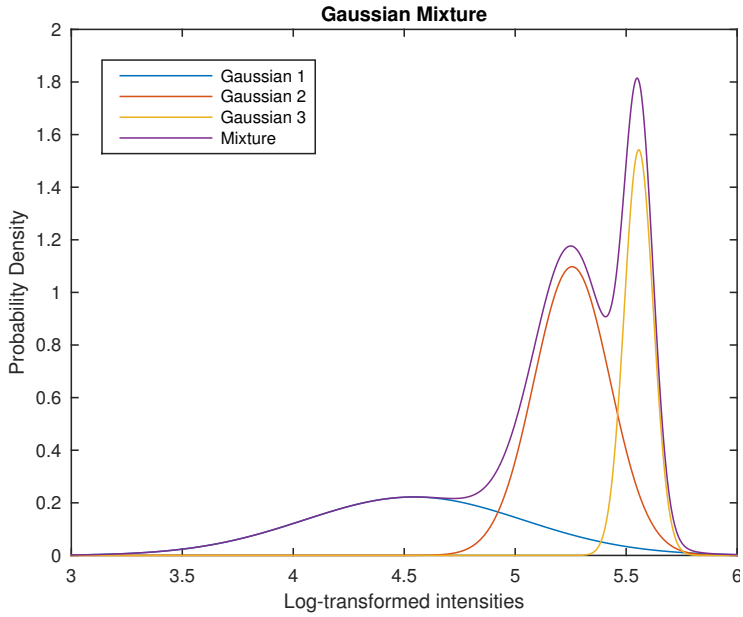


Figure 5.3: A Gaussian mixture with predetermined parameters for a log-transformed 3T MRI and $L = 3$, corresponding to CSF, GM and WM respectively in brain MRI.

Finally, by summing over the probabilities for observing $p(\boldsymbol{d}|\boldsymbol{l}, \boldsymbol{\theta}_d)$ for each label, we obtain the likelihood function for $\boldsymbol{d}$ given all model parameters $\boldsymbol{\theta}$

$$p(\boldsymbol{d}|\boldsymbol{\theta}) = \sum_{l=1}^{L} p(\boldsymbol{d}|\boldsymbol{l}, \boldsymbol{\theta}_d) p(\boldsymbol{l}|\boldsymbol{\theta}_l) \tag{5.14}$$

$$= \prod_{i=1}^{N} \sum_{l=1}^{L} \mathcal{N}(d_i|\mu_l, \sigma_l^2)\pi_l, \tag{5.15}$$

with parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_d^T, \boldsymbol{\theta}_l^T)^T = (\mu_1, \ldots, \mu_L, \sigma_1^2, \ldots, \sigma_L^2, \pi_1, \ldots, \pi_L)^T$. Here, π_l is also referred to as the mixing coefficient, as it weighs each Gaussian in the linear superposition of Gaussians such that unit area under the curve of the PDF is preserved, i.e., respecting Equation 5.1.

Since the model encodes no spatial information about the distribution of the intensities generated in the image, it can be seen as a way to approximate the histogram. A mixture composed of $L = 3$ Gaussians with predetermined parameters for a 3T log-transformed MRI have been illustrated in Figure 5.3.

5.1.4 Modeling MR Images with a Bias Field

At this point we are ready to discuss how we can include the bias field effect in a generative model. First, we assume that the bias field effect that we try to model is multiplicative. This assumption is actually only an approximation, as the field is known to be discontinuous across tissue borders. Regardless, it leads to a simplified model that has shown to work well when the goal is to correct data for the purpose of segmentation [Styner and Van Leemput 2004].

Using $\hat{d}$ to denote the observed intensity of a voxel, $\hat{b}$ to denote the effect due to the bias field, $\hat{u}$ to denote the underlying "true" intensity and finally $\hat{n}$ to denote noise due to acquisition, we have

$$\hat{d} = \hat{b}\hat{u} + \hat{n}. \tag{5.16}$$

To simplify the model, we further assume that we can ignore the effects of $\hat{n}$ and log-transform the data $d = \log(\hat{d})$, as it simplifies the model [W. M. Wells et al. 1996; Van Leemput et al. 1999a; Zhang et al. 2001]. Using $\boldsymbol{d} = (d_i, \ldots, d_N)$ to denote all log-transformed voxel intensities of an MR image, and similarly $\boldsymbol{b} = (b_1, \ldots, b_N)^T$ to denote the corresponding (log-transformed) gains due to the bias field, we have

$$\boldsymbol{d} = \boldsymbol{u} + \boldsymbol{b}, \tag{5.17}$$

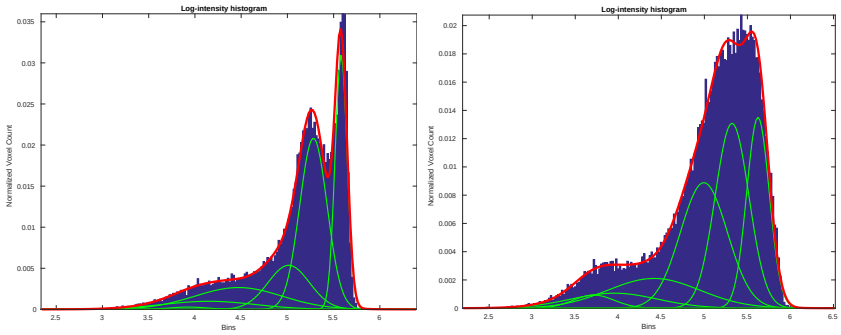
where $\boldsymbol{u} = (u_1, \ldots, u_N)^T$ are the intensities of the "true", underlying intensities that are not affected by the bias field. It is worth mentioning that other models have been proposed on how to best model the bias field effect, of which [Vovk et al. 2007] provides an excellent overview.

For now, we assume that the shape of the bias field $\boldsymbol{b}$ is known. Given mixture model parameters $\boldsymbol{\theta}_d = (\mu_1, \ldots, \mu_L, \sigma_1^2, \ldots, \sigma_L^2, \pi_1, \ldots, \pi_L)^T$, we have

$$p(\boldsymbol{d}|\boldsymbol{\theta}_d) = \prod_{i=1}^{N} \sum_{l=1}^{L} \mathcal{N}(d_i - b_i|\mu_l, \sigma_l^2)\pi_l. \tag{5.18}$$

Note that the mixture model describes the underlying "true" distribution of $\boldsymbol{u}$, and the *forward* model is that of Eq 5.17: the data we observe is generated by sampling our mixture model to obtain $\boldsymbol{u}$ (a), and then adding the bias $\boldsymbol{b}$. Figure 5.4 (approximately) illustrate how the histogram of the generated data would appear at each step.

It should be mentioned that the histograms were not obtained by sampling, but rather by bias field correcting a real image, but regardless, the example is analogous. It can be seen how adding the bias field results in a histogram that



(a) (b)

Figure 5.4: The (log-ingensity) histogram of MRI data after (a) and before (b) bias field correction using a mixture model of $L = 6$ Gaussians, which have been overlaid (green lines) together with their sum (red line).The histograms are approximately analogues to first generating data $\boldsymbol{u}$, and then adding bias $\boldsymbol{b}$ (b).

becomes wider. This is the reason why bias field correction methods (e.g., the N3 algorithm) claim to *sharpen* the histogram; it is the obvious consequence of removing the bias.

It's important to realize that an image generated by the model will only resemble that of the observed one in terms of its statistical content (the histogram), as the voxel intensities are modeled to be independent on their spatial position. Spatial information about voxel intensities is typically encoded by defining a distribution using a probabilistic atlas, such that each voxel has its own set of probabilities for belonging to each label, e.g., WM, GM and CSF in the brain. Whereas a model utilizing a probabilistic atlas will generate more realistic images, this simpler, purely intensity-based model has the advantage that it is not only limited to images where label priors are available, which is typically the case for brain MRI, but not e.g., the abdomen.

Modeling a Smooth Bias Field

So far, we have not discussed how to model the shape of the bias field $\boldsymbol{b}$. As we're working under the assumption of a smooth and slowly varying field over the image, one way to model this is to use a linear combination of smooth basis functions, such as splines, low order polynomials, or cosine functions. Using this bias field model, we have for M basis functions $\boldsymbol{\phi} = (\phi_{i,1}, \ldots, \phi_{i,M})^T$ evaluated

at voxel i and with coefficients $\boldsymbol{c} = (c_1, \ldots, c_M)^T$

$$b_i = \sum_{m=1}^{M} c_m \phi_{i,m}, \tag{5.19}$$

or, in matrix notation,

$$\boldsymbol{b} = \boldsymbol{\Phi c}. \tag{5.20}$$

The bias field coefficients $\boldsymbol{c}$ are assumed to be generated by a prior distribution $p(\boldsymbol{c})$. Consequently, in its most general form, the generative bias field model is composed of both parameters for the mixture model as well as the bias field coefficients $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_L, \sigma_1^2, \ldots, \sigma_L^2, \pi_1, \ldots, \pi_L, c_1, \ldots, c_M)$.

Practically, $\boldsymbol{\Phi}$ is computed by evaluating the one-dimensional basis functions of choice along each of the three dimensions of the MR image. Assuming an MR image of dimensions $N_x \times N_y \times N_z$ and a total of $M = M_x M_y M_z$ basis functions where M_x, M_y and M_z are the number of basis functions evaluated along each of the three dimensions respectively, the matrix $\boldsymbol{\Phi}$ contains the full number of basis functions evaluated for all voxels. $\boldsymbol{\Phi}$ is obtained by the Kronecker product

$$\boldsymbol{\Phi} = \boldsymbol{\Phi}_x \otimes \boldsymbol{\Phi}_y \otimes \boldsymbol{\Phi}_z, \tag{5.21}$$

with e.g. the matrix $\boldsymbol{\Phi}_x$ defined as

$$\boldsymbol{\Phi}_x = \begin{pmatrix} \phi_{1,1} & \phi_{1,2} & \cdots & \phi_{1,M_x} \\ \phi_{2,1} & \phi_{2,2} & \cdots & \phi_{2,M_x} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{N_x,1} & \phi_{N_x,2} & \cdots & \phi_{N_x,M_x} \end{pmatrix}. \tag{5.22}$$

Basis functions are generally chosen based on preference or because they have desirable properties. Theory does not suggest that one type of basis function leads to better bias field correction than another, as the model only assumes smoothness. All basis functions that will be described in the following meet this assumption.

Cosines have the nice feature that they have global support, meaning that they span the entire image. This prevents against numerical issues due to estimation of the basis function coefficients in areas of the brain which has been masked out. Cosine basis functions are computed according to (e.g, for cosine basis function m along the dimension of x):

$$\phi_{x,m} = \omega(m) \cos\left(\frac{\pi}{2X}(2x-1)(m-1)\right), \quad m = 1, 2, \ldots, M \tag{5.23}$$

with

$$\omega(m) = \begin{cases} \frac{1}{\sqrt{X}}, & m = 1 \\ \sqrt{\frac{2}{X}}, & 2 \leq m \leq M, \end{cases}$$

where M is the number of basis functions, x denotes the position of $i = 1, 2, \ldots, N_x$ voxels along the dimension of x, and X is the position of voxel N_x.

Splines have local support, meaning that they may be more adaptable to variations in the image. Using one of several equally valid definitions, cubic b-splines are computed according to (e.g., for spline m along the dimension of x):

$$\phi_{x,m} = \sum_{s=0}^{4} \frac{-1^s}{h^3} \binom{4}{s} \left[x - \lambda_{m-s}\right]^3 \omega\left(x - \lambda_{m-s}\right), \tag{5.24}$$

with

$$\omega(k) = \begin{cases} 1, & k \geq 0 \\ 0, & k < 0, \end{cases}$$

where λ_{m-s} is one of $M_x + 3$ knot locations and h is the distance between them. This is the spline scheme used in [Sled et al. 1998].

Because the support of a given spline might not cover a sufficient number of foreground voxels to perform the estimation of its coefficient, regularization may be necessary.

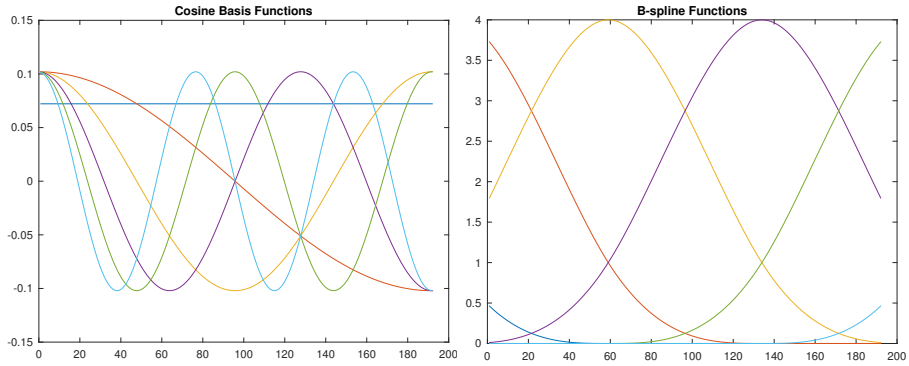
Polynomials were not explored in this thesis, but have previously been used in e.g.,[Van Leemput et al. 1999a].

Cubic b-splines were predominantly explored throughout this work, in order to maintain comparability with the N3 algorithm. Cosine basis functions are described here for completeness, in part due to the nice feature of full support with makes them a complementary choice next to cubic b splines, and in part because they're used in SPM[Ashburner and Friston 2005]. Examples of cosine and cubic spline basis functions have been illustrated in Figure5.5

5.1.5 A Unified Model for Bias Field Correction

By expressing the model for bias field correction in terms of the joint probability

$$p(\boldsymbol{d}, \boldsymbol{l}, \boldsymbol{\theta}, \boldsymbol{c}) = p(\boldsymbol{d} - \boldsymbol{\Phi c} | \boldsymbol{l}, \boldsymbol{\theta}_d) p(\boldsymbol{l} | \boldsymbol{\theta}_l) p(\boldsymbol{\theta}_d) p(\boldsymbol{\theta}_l) p(\boldsymbol{c}), \tag{5.25}$$



(a) (b)

Figure 5.5: Cosine basis (a), and B-spline (b) functions in one dimension.

$\boldsymbol{\theta_l}$	$\sim$	$p(\boldsymbol{\theta}_l)$
$\boldsymbol{l}$	$\sim$	$p(\boldsymbol{l}\|\boldsymbol{\theta}_l) = \prod_{i=1}^{N} p(l_i\|\boldsymbol{\theta}_l)$
$\boldsymbol{\theta}_d$	$\sim$	$p(\boldsymbol{\theta}_d)$
$\boldsymbol{u}$	$\sim$	$p(\boldsymbol{u}\|\boldsymbol{l}, \boldsymbol{\theta}_d) = \prod_{i=1}^{N} p(u_i\|l_i, \boldsymbol{\theta}_d)$
$\boldsymbol{c}$	$\sim$	$p(\boldsymbol{c})$
$\boldsymbol{b}$	$=$	$\boldsymbol{\Phi c}$
$\boldsymbol{d}$	$=$	$\boldsymbol{u} + \boldsymbol{b}$

Table 5.1: Generative model of bias field corrupted MRI data.

where $p(\boldsymbol{d} - \boldsymbol{\Phi c}|\boldsymbol{l}, \boldsymbol{\theta}_d)$ is the probability for observing the true underlying image $\boldsymbol{u} = \boldsymbol{d} - \boldsymbol{\Phi c}$ given all model parameters and labels, $p(\boldsymbol{l}|\boldsymbol{\theta}_l)$ the probability of observing a label given the label parameters, and finally $p(\boldsymbol{\theta}_d)$, $p(\boldsymbol{\theta}_l)$, and $p(\boldsymbol{c})$ are prior distributions on the parameters themselves, we have a very powerful generative framework in place which allows us to create much more complicated models than before. This is done simply by choosing appropriate distributions – or combinations thereof – for each of the involved PDFs.

Figure 5.6 shows a directed graph that describes the full generative bias field correction framework, which serves as the basis for the models presented in paper C. Circles illustrate latent variables, whereas shaded circles are those that are observed. In this thesis, we only consider uniform priors of the form $p(\boldsymbol{\theta}_d) \propto 1$ and $p(\boldsymbol{\theta}_l) \propto 1$. This is a typical approach, but the significance is that we do not *have* to represent them this way. It follows that the ability to express everything as distributions and then modify as needed by choosing appropriate PDFs, is one of the key advantages of using generative models. The full list of distributions have been described in table 5.1.

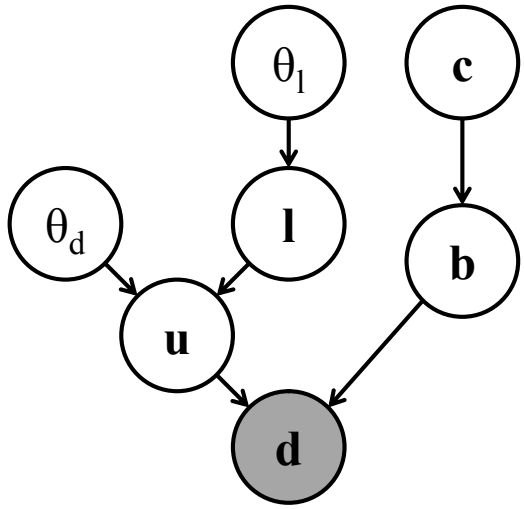


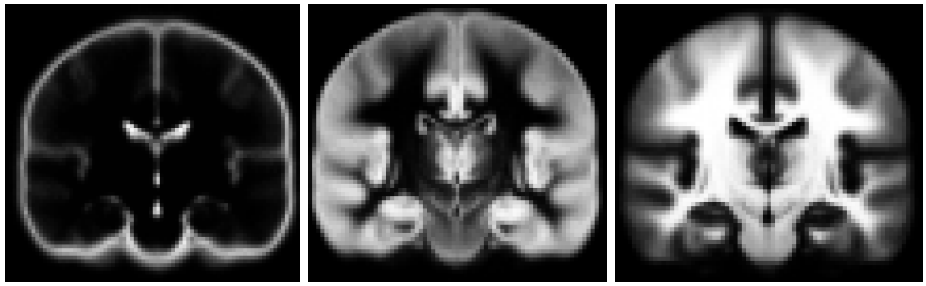
Figure 5.6: A directed graph showing the full generative model for bias field correction. The shaded circle illustrate the observed variable (bias field corrupted data), whereas the rest are latent variables.

5.1.6 Useful Priors

The prior distributions of $\boldsymbol{l}$, $\boldsymbol{\theta}_l$, $\boldsymbol{\theta}_d$ and $\boldsymbol{c}$ yield a very powerful mechanism for controlling/informing our model. In the following, we briefly discuss these priors, in particular on the distribution of the bias field coefficients and labels, as they are of particular relevance in bias field correction.

Label Priors

As previously described in Section 5.1.3, we do not make any a-priori assumptions about the distribution of labels when we let the label prior equal the relative frequency for a label to appear in a voxel, on average, $p(l_i|\boldsymbol{\theta}_l) = \pi_l$. In this case, we are modeling the image solely in terms of its histogram. Typically, this is relevant when we do not have an anatomical atlas available. For example, in brain MRI a very simple model using a Gaussian mixture of $L = 3$ labels can be used to represent WM, GM and CSF, as was shown in Figure 5.3. It is *possible* to further impose a label prior $p(\boldsymbol{l}|\boldsymbol{\theta}_l)$ on this model by means of a probabilistic tissue atlas, and this would most likely improve bias field correction, but it is not *required*. Furthermore, since any distribution can be expressed by a super-



(a) (b) (c)

Figure 5.7: A probabilistic tissue atlas of CSF (a), GM (b) and WM (c).

position of (more) Gaussians, we can model the same data by setting $L = 200$ and choosing appropriate parameter values $\boldsymbol{\theta}_d = (\mu_1, \ldots, \mu_L, \sigma_1^2, \ldots, \sigma_L^2)^T$. This is how data is generated in the model behind the N3 algorithm [Larsen et al. 2014], where l no longer has any anatomically meaningful interpretion – it is purely abstract.

Anatomical atlases are of particular interest in the context of labels. We can inform the model about label probabilities using a discrete distribution (since we have a finite number of labels)

$$p(l_i|\boldsymbol{\theta}) = A_{il}, \tag{5.26}$$

Where A_{il} is the probability for a voxel i to belong to label l, satisfying $0 \leq A_{il} \leq 1$ and $\sum_{l=1}^{L} A_{il} = 1$. As an example, Fig 5.7 illustrates a probabilistic tissue atlas of observing each of the three labels: WM, GM and CSF. The atlas is from SPM, and used in the old_segment method of the software, based on [Ashburner and Friston 2005].

In cases where we use such atlases, it is sometimes of interest to model $\boldsymbol{u}$ using separate mixture models for each label l, each composed of its own superposition of K_l gaussians:

$$p(\boldsymbol{u}|l, \boldsymbol{\theta}_d,) = \prod_{i=1}^{N} \sum_{k=1}^{K_l} \mathcal{N}(u_i|\mu_{lk}, \sigma_{lk}^2)\pi_{lk}. \tag{5.27}$$

This addresses concerns where the intensities of each tissue or structure represented by label l does not follow a single Gaussian distribution. Typically, a few Gaussians (two or three) in each mixture is enough to model each label well.

Mixture Model Parameter Priors

Regularization on the values that the parameters can take can be imposed by choosing appropriate distributions for $p(\boldsymbol{\theta}_d)$. As mentioned we generally assume $p(\boldsymbol{\theta}_d) \propto 1$. Although not a topic of focus in this work, for computational reasons it may be of (particular) interest to regularize the Gaussian variance.

Bias Field Coefficient Priors

While cubic B-spline basis functions may require regularization, explicit regularization in the bias field prior can also protect against basis functions that are too flexible. A convenient prior on the bias field coefficients is

$$p(\boldsymbol{c}) \propto \exp(-\lambda \boldsymbol{c}^T \boldsymbol{\Psi} \boldsymbol{c}), \tag{5.28}$$

where $\boldsymbol{\Psi}$ is a positive semi-definite regularization matrix. Some works (e.g., [Van Leemput et al. 1999a]) use a uniform prior instead, since the polynomial basis functions by themselves are smooth and require no regularization. [W. M. Wells et al. 1996] use the identity matrix for the basis functions, and depends solely on the expression for $\boldsymbol{\Psi}$ to obtain smooth functions.

Of particular interest to us is an $M \times M$ regularization matrix that penalizes the bending energy of a basis function ϕ of the form

$$J_p(\phi) = \frac{1}{V} \int_{\mathbb{R}^p} \sum_{i=1}^{P} \sum_{j=1}^{P} \left[\frac{\partial^2 \phi}{\partial u_i \partial u_j} \right]^2 d\boldsymbol{u}, \tag{5.29}$$

where ϕ is the basis function that is evaluated, $\boldsymbol{u} = [x, y, z]$ is the position the function is evaluated at, P is the dimensionality of the image data, and V is the volume of the region of interest.

Cubic b-splines: the bending energy regularization matrix (three-dimensional case $P = 3$) is defined as

$$\boldsymbol{\Psi} = \sum_{\substack{\alpha_x, \alpha_y, \alpha_z \geq 0 \\ \alpha_x + \alpha_y + \alpha_z = 2}} \frac{2}{\alpha_x! \alpha_y! \alpha_z!} \boldsymbol{\Psi}_x^{(\alpha_x)} \otimes \boldsymbol{\Psi}_y^{(\alpha_y)} \otimes \boldsymbol{\Psi}_z^{(\alpha_z)}, \tag{5.30}$$

with elements e.g., for $\boldsymbol{\Psi}_x^{(\alpha_x)}$

$$\psi_{i,j}^{(\alpha)} = \frac{1}{V} \int_D = \phi_i^{(\alpha)}(x) \phi_j^{(\alpha)}(x) \, \mathrm{d}x, \tag{5.31}$$

where e.g., $\phi_i^{(\alpha)}(x)$ denotes the α'th derivative of the i'th basis function evaluated at x. The region D can be expressed in terms of the knot locations

$$D = \left[\lambda_0^{(x)}, \lambda_{M_x-3}^{(x)}\right] \times \left[\lambda_0^{(y)}, \lambda_{M_y-3}^{(y)}\right] \times \left[\lambda_0^{(z)}, \lambda_{M_z-3}^{(z)}\right], \tag{5.32}$$

and V is the volume of D. This is the regularization employed on the splines in the N3 algorithm. The definitions used here were presented in [Sled et al. 1998], with minor modifications in notation. [Shackleford et al. 2012] presents analytical derivations that can be used for practical implementations of Equation 5.31.

5.2 Maximum A Posteriori Probability (MAP) Model Parameter Estimation

We now turn our attention to the cases where we need to estimate the model parameters from the data, which is typically the problem we need to solve. Specifically, we look for the maximum a posteriori parameters, i..e, those that are most probable to have generated the data we observe.

5.2.1 Bayes' Theorem

Bayes' Theorem states

$$p(\boldsymbol{\theta}|\boldsymbol{d}) = \frac{p(\boldsymbol{d}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{d})}, \tag{5.33}$$

and consequently

$$p(\boldsymbol{\theta}|\boldsymbol{d}) \propto p(\boldsymbol{d}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{5.34}$$

This forms the basis for estimating the parameters of any generative model.

The maximum a posteriori (MAP) parameters are given by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[p(\boldsymbol{\theta}|\boldsymbol{d})\right] = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[p(\boldsymbol{d}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right] \tag{5.35}$$

$$= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\log p(\boldsymbol{\theta}|\boldsymbol{d})\right] = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\log p(\boldsymbol{d}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\right], \tag{5.36}$$

where we refer to $\log p(\boldsymbol{d}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$ as the *objective* function. This problem is then maximized choosing an appropriate optimizer. One such is the Expectation Maximization (EM) algorithm [Dempster et al. 1977; Minka 1998] which is very suitable for models that encompass labels that are unobserved. The reason for

the log-tranformation is mostly practical, as it simplifies mathematical analysis and protects against numerical issues in implementations. The transformation is valid, because the logarithm is a monotonically increasing function of its argument. Somewhat intuitively described, this means that the transformation will not change the "peaks of the parameter landscape", which we need to traverse when we estimate the optimal parameters.

5.2.2 Expectation Maximization

EM iteratively builds a lower bound $\varphi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ of the objective function that touches it at the current estimate $\tilde{\boldsymbol{\theta}}$ of the model parameters (E step), and subsequently improves $\varphi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ with respect to the parameters (M step). This procedure automatically guarantees to increase the value of the objective function at each iteration, which is a highly desirable property in an optimizer.

The lower bound is formulated by means of Jensen's inequality. The inequality states

$$\log\left(\sum_{l=1}^{L} w_l x_l\right) \geq \sum_{l=1}^{L} w_l \log(x_l), \tag{5.37}$$

which satisfies $w_l \geq 0$ and $\sum_{l=1}^{L} w_l = 1$ for any value of x_l. Consequently, we have

$$\begin{aligned}
\log p(\boldsymbol{d}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) &= \log\left(\prod_{i=1}^{N}\left[\sum_{l=1}^{L} p(d_i - b_i|l,\boldsymbol{\theta})p(l|\boldsymbol{\theta})\right]\right) + \log p(\boldsymbol{\theta}) \\
&= \sum_{i=1}^{N} \log\left[\sum_{l=1}^{L} p(d_i - b_i|l,\boldsymbol{\theta})p(l_i|\boldsymbol{\theta})\right] + \log p(\boldsymbol{\theta}) \\
&= \sum_{i=1}^{N} \log\left[\sum_{l=1}^{L} w_i^l \left(\frac{p(d_i - b_i|l,\boldsymbol{\theta})p(l|\boldsymbol{\theta})}{w_i^l}\right)\right] + \log p(\boldsymbol{\theta}) \\
&\geq \underbrace{\sum_{i=1}^{N}\left[\sum_{l=1}^{L} w_i^l \log\left(\frac{p(d_i - b_i|l,\boldsymbol{\theta})p(l|\boldsymbol{\theta})}{w_i^l}\right)\right] + \log p(\boldsymbol{\theta})}_{\varphi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})}.
\end{aligned} \tag{5.38}$$

The lower bound must satisfy two conditions to be valid. The first condition is that the it should touch the objective function at the current parameter estimate

$$\varphi(\tilde{\boldsymbol{\theta}}|\tilde{\boldsymbol{\theta}}) = \log p(\boldsymbol{d}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}), \tag{5.39}$$

and the second, that it should *never* exceed the objective function

$$\varphi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}) \leq \log p(\boldsymbol{d}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}), \forall \boldsymbol{\theta}. \tag{5.40}$$

The second condition is already met, given the use of Jensen's inequality. It can be shown [Minka 1998] that the first condition is met by choosing the weights w_l^i such that they describe the posterior probability $p(l|d_i, \tilde{\boldsymbol{\theta}})$ for a voxel to belong to any of the l Gaussians (we *soft-assign* the voxels):

$$w_l^i = p(l|d_i, \tilde{\boldsymbol{\theta}}) = \frac{p(d_i - \tilde{b}_i|l, \tilde{\boldsymbol{\theta}})p(l|\tilde{\boldsymbol{\theta}})}{\sum_{l'=1}^{L} p(d_i - \tilde{b}_i|l, \tilde{\boldsymbol{\theta}})p(l|\tilde{\boldsymbol{\theta}})}. \tag{5.41}$$

Computing the posterior consequently constructs the lower bound (E-Step). Given the partial derivatives of the lower bound with respect to each of the involved parameters $\tilde{\boldsymbol{\theta}}$ (still assuming that the bias field coefficients $\boldsymbol{c}$ are known), and setting these to zero, we obtain corresponding update equations that can be used to maximize the lower bound (M-Step).

The objective function is then iteratively maximized by alternating between the E-step and M-step. This process has been illustrated for the objective function in Figure 5.8 [Van Leemput and Puonti 2015] (using a uniform prior $p(\boldsymbol{\theta}) \propto 1$ for the parameters, as this simplifies the example but does not change its validity).

5.3 Building a Bias Field Correction Algorithm

In the following we consider just one bias field correction model to exemplify how a bias field correction algorithm is built using GEM for parameter estimation. More advanced models have been covered in paper C. The process can be broken down into the following four steps

1. Define model.
2. Derive lower bound and expression for weights (posterior label probabilities).
3. Derive expressions the involved parameter updates.
4. Defining the algorithm.

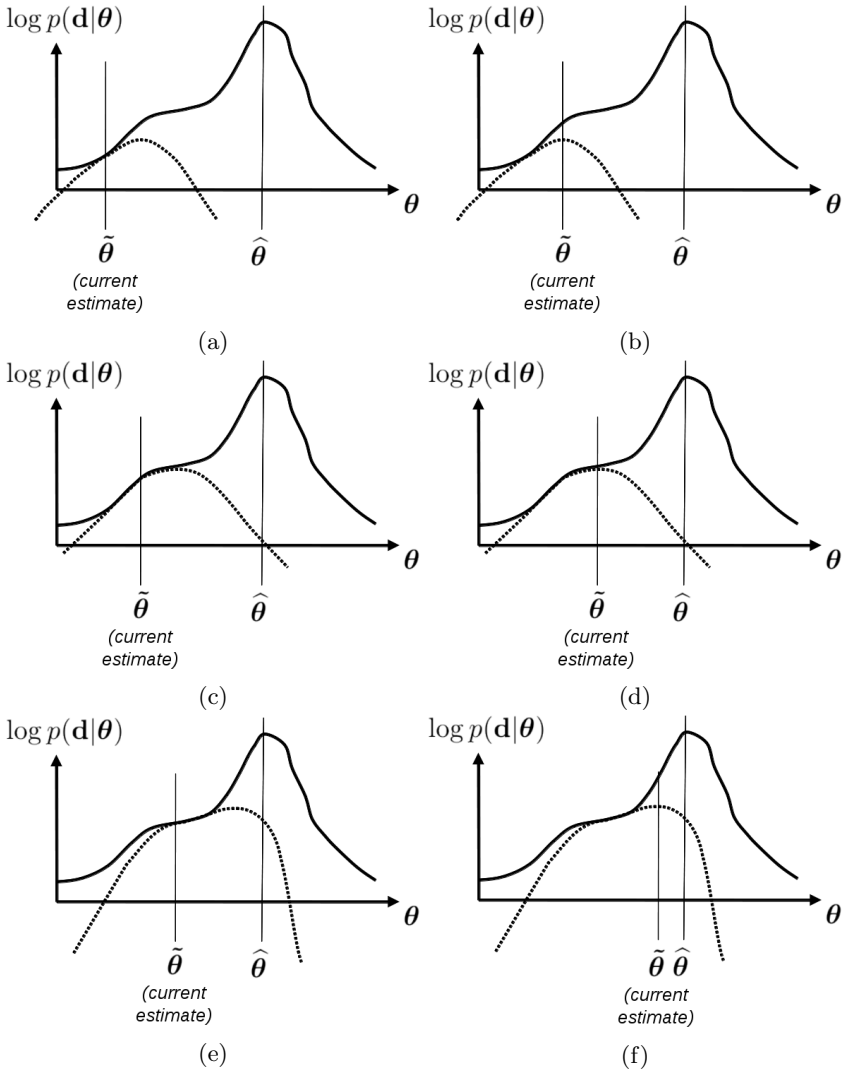


Figure 5.8: Estimation of the maximum a-posteriori parameters $\hat{\boldsymbol{\theta}}$ using the Expectation Maximization algorithm is done by alternating between constructing the lower bound to the objective function in the E-step (a, c, e), and then estimating the parameters that maximize the lower bound in the M-step (b, d, f). The process alternates until the parameter estimates have converged. The objective function is represented by a full line, and the lower bound with a broken line. Here we assumed $p(\boldsymbol{\theta}) \propto 1$, which means that we only need to estimate the *maximum likelihood* (ML) parameters of the likelihood function $p(\boldsymbol{d}|\boldsymbol{\theta})$.

5.3.1 Defining the model

We consider the general model where all parameters are unknown and needs to be learned, i.e., $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_L, \sigma_1^2, \ldots, \sigma_L^2, \pi_1, \ldots, \pi_L, c_1, \ldots, c_M)$. We use a cubic B-spline basis of M functions to model the bias field $\boldsymbol{b} = \boldsymbol{\Phi c}$, and we penalize curvature of the bias field using $p(\boldsymbol{c}) \propto \exp(-\lambda \boldsymbol{c}^T \boldsymbol{\Psi c})$. Keeping the label and mixture parameter priors uniform $p(\boldsymbol{\theta}_d) \propto 1$ and $p(\boldsymbol{\theta}_l) \propto 1$, the objective function is (with constant terms referred to in the following by k)

$$\log p(\boldsymbol{\theta}|\boldsymbol{d}) = \sum_{i=1}^{N} \log \left(\sum_{l=1}^{L} \mathcal{N}(d_i - b_i|\mu_l, \sigma_l^2)\pi_l \right) - \lambda \boldsymbol{c}^T \boldsymbol{\Psi c} + k. \tag{5.42}$$

5.3.2 Deriving the Lower Bound

The model yields the following lower bound

$$\varphi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}) = \sum_{i=1}^{N} \left[\sum_{l=1}^{L} w_l^i \log \left(\frac{\mathcal{N}(d_i - b_i|\mu_l, \sigma_l^2)\pi_l}{w_l^i} \right) \right] - \lambda \boldsymbol{c}^T \boldsymbol{\Psi c} + k, \tag{5.43}$$

with weights (label probabilities) in each voxel:

$$w_l^i = \frac{\mathcal{N}\left(d_i - b_i|\tilde{\mu}_l, \tilde{\sigma}_l^2\right)\tilde{\pi}_l}{\sum_{l'=1}^{L} \mathcal{N}\left(d_i - b_i|\tilde{\mu}_{l'}, \tilde{\sigma}_{l'}^2\right)\tilde{\pi}_{l'}}. \tag{5.44}$$

Writing out and rearranging the terms of $\varphi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ yields

$$\begin{aligned} \varphi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}) = & -\frac{1}{2} \sum_{l=1}^{L} \left[\frac{1}{\sigma_l^2} \sum_{i=1}^{N} w_l^i (d_i - \mu_l - b_i)^2 + \log(\sigma_l^2) \sum_{i=1}^{N} w_l^i \right] \\ & + \sum_{l=1}^{L} \left[\log(\pi_l) \sum_{i=1}^{N} w_l^i \right] \\ & - \sum_{l=1}^{L} \left[\sum_{i=1}^{N} w_l^i \log(w_l^i) \right] \\ & - \lambda \boldsymbol{c}^T \boldsymbol{\Psi c} \\ & - \frac{N}{2} \log(2\pi). \end{aligned} \tag{5.45}$$

When parameters are interdependent, as in the case of the bias field coefficients $\boldsymbol{c}$ and the mixture model parameters, optimization is no longer possible using the EM algorithm. However, optimization of one with respect to a given set of the other is closed form. As such, optimal parameters can be estimated by fixing the coefficients and updating the mixture parameters, and vice versa.

5.3.3 Deriving the Parameter Updates

We obtain the following parameter updates:

Label mean (μ_l):

$$\frac{\mathrm{d}\varphi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})}{\mathrm{d}\mu_l} = -\frac{1}{\sigma_l^2}\sum_{i=1}^{N} w_l^i(d_i - \mu_l - b_i) = -\frac{1}{\sigma_l^2}\left(\sum_{i=1}^{N} w_l^i(d_i - b_i) + \mu_l \sum_{i=1}^{N} w_l^i\right),$$

which yields the following update

$$\mu_l \leftarrow \frac{\sum_{i=1}^{N} w_l^i(d_i - b_i)}{\sum_{i=1}^{N} w_l^i}. \tag{5.46}$$

Label variance (σ_l^2):

$$\frac{\mathrm{d}\varphi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})}{\mathrm{d}\sigma_l^2} = -\frac{1}{2\sigma_l^2}\left(\sum_{i=1}^{N} w_l^i + \frac{1}{\sigma_l^2}\sum_{i=1}^{N} w_l^i(d_i - \mu_l - b_i)^2\right),$$

thereby obtaining the update

$$\sigma_l^2 \leftarrow \frac{\sum_{i=1}^{N} w_l^i(d_i - \mu_l - b_i)^2}{\sum_{i=1}^{N} w_l^i}. \tag{5.47}$$

Label Frequency (π_l):

For $\boldsymbol{\pi}$ we have the condition that $\sum_{l=1}^{L} \pi_l = 1$. Using a Lagrange multiplier, we have

$$\Lambda(\pi_l, \lambda) = \varphi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}) + \lambda\left(\sum_{l=1}^{L} \pi_l - 1\right) = \sum_{l=1}^{L} \log(\pi_l)\left(\sum_{i=1}^{N} w_l^i\right) + \lambda\left(\sum_{l=1}^{L} \pi_l - 1\right) + k.$$

Taking partial derivatives yields

$$\frac{\mathrm{d}\Lambda}{\mathrm{d}\pi_l} = \frac{1}{\pi_l}\sum_{i=1}^{N} w_l^i + \lambda, \quad \frac{\mathrm{d}\Lambda}{\mathrm{d}\lambda} = \sum_{l=1}^{L} \pi_l - 1.$$

Setting each partial derivative to zero yields

$$\sum_{l=1}^{L} \pi_l = 1, \quad \pi_l = -\frac{1}{\lambda} \sum_{i=1}^{N} w_l^i.$$

To determine the value for λ that fulfills the condition $\sum_l \pi_l = 1$ we substitute the second expression into the first and obtain

$$\lambda = -\sum_{i=1}^{N} \sum_{l=1}^{L} w_l^i = -N$$

with $\sum_l w_l^i = 1, \forall l$. Substituting this expression into the previously obtained expression for π_l we obtain the following parameter update

$$\pi_l \leftarrow \frac{\sum_{i=1}^{N} w_l^i}{N}. \tag{5.48}$$

We see that the label frequency is a sum over the posterior label probabilities in all voxels – we estimate the prior probability for observing the label l.

Bias Field Coefficients ($\boldsymbol{c}$)

We have

$$\varphi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}) = -\frac{1}{2} \sum_{l=1}^{L} \frac{1}{\sigma_l^2} \sum_i w_l^i \left(d_i - \mu_l - \sum_m c_m \phi_m^i \right)^2 - \lambda \boldsymbol{c}^\top \boldsymbol{\Psi} \boldsymbol{c} + k,$$

which can be rewritten into the following regression problem[1]

$$\varphi = -\frac{1}{2} \parallel \boldsymbol{S}^{\frac{1}{2}} (\boldsymbol{r} - \boldsymbol{\Phi} \boldsymbol{c}) \parallel^2 - \lambda \boldsymbol{c}^\top \boldsymbol{\Psi} \boldsymbol{c} + k,$$

where we defined

$$s_l^i = \frac{w_l^i}{\sigma_l^2}, \quad s_i = \sum_{l=1}^{L} s_l^i, \quad \boldsymbol{S} = \mathrm{diag}(s_i), \quad \bar{d}_i = \frac{\sum_{l=1}^{L} s_l^i \mu_l}{\sum_{l'=1}^{L} s_{l'}^i}, \quad \boldsymbol{r} = \boldsymbol{d} - \bar{\boldsymbol{d}}.$$

Taking the partial derivative yields

$$\frac{\mathrm{d}\varphi}{\mathrm{d}\boldsymbol{c}} = -\boldsymbol{\Phi}^\top \boldsymbol{S} \boldsymbol{\Phi} \boldsymbol{c} + \boldsymbol{\Phi}^\top \boldsymbol{S} \boldsymbol{r} - 2\lambda \boldsymbol{\Psi} \boldsymbol{c},$$

which results in the following update (a regularized least-squares fit)

$$\boldsymbol{c} \leftarrow (\boldsymbol{\Phi}^\top \boldsymbol{S} \boldsymbol{\Phi} + 2\lambda \boldsymbol{\Psi})^{-1} \boldsymbol{\Phi}^\top \boldsymbol{S} \boldsymbol{r}. \tag{5.49}$$

[1] The rearrangement of terms is not entirely trivial, but omitted for brevity.

An estimate of the bias field is then obtained with $\tilde{\boldsymbol{b}} = \boldsymbol{\Phi c}$. The equations reveal that the bias field estimate is obtained by smoothing the residual intensities $\boldsymbol{r}$, which are the difference between the expected "true" voxel intensities $\bar{\boldsymbol{d}}$ produced by the model and the observations $\boldsymbol{d}$. We further see that voxels that belong to Gaussians with wide variance (represented by the matrix $\boldsymbol{S}$), have less weight in the regularized least-squares fit that yields the bias field coefficients.

5.3.4 Defining the Algorithm

By updating only some or all of the parameters once and in the M-step, and then recomputing the E step, a *generalized* Expectation Maximization (GEM) algorithm is obtained. The algorithm still guarantees an increase in the objective function at every iteration - the lower bound is just never fully maximized before the posterior is recomputed.

We assume that the most accurate bias field estimates are obtained when the mixture model fits the data as closely as possible; the more accurate we model the true underlying intensities $\boldsymbol{u} = \boldsymbol{d} - \boldsymbol{b}$, the more accurate our expectations and consequent bias field estimates become. A single bias field coefficient update given these residuals will in turn affect our estimate of the distribution of $\boldsymbol{u}$, which means that we need to refit the mixture model.

Following these guidelines, we design the algorithm:

1. Initialize the algorithm (set parameter estimates to some initial values, e.g., the bias field is assumed initially flat $\boldsymbol{b} = 0$, Gaussian means are equidistantly spaced over the span of the data, some appropriate values for variance are chosen, e.g., the squared mean spacing and label frequencies are set to be equal $\pi_l = 1/l$).
2. Fit the mixture model to the current estimate of the "true" data $\boldsymbol{u} = \boldsymbol{d} - \boldsymbol{b}$ by alternating between updating the posterior and the mixture model parameters until an appropriate convergence has been obtained.
3. Fit the bias field coefficients once.
4. Repeat 2-3 until global convergence.

Convergence of the parameter estimation can be determined by e.g., evaluating the absolute or relative change in the objective function per iteration, or alternatively by evaluating the change in the standard deviation of two consecutive

bias field estimates subtracted (which is related to the change in the objective function, as the bias field estimates depend on the parameters).

This algorithm design has been used throughout the work performed in this study, and it has proven to yield good results. Ultimately, the algorithm simply defines one way to traverse parameter space in order to reach a maximum of the objective function. This means that the order and arrangement of the parameter updates may affect both what maximum value of the objective function we find (we're never guaranteed to reach the global maximum), and also how many updates we need to perform.

Of particular interest is the variance, which can be interpreted both as how certain the model is about the intensities it generates given some label, and also as a way to control the speed of the algorithm. The greater the variance of each label is, the larger the step size between iterations becomes, i.e., parameter values differ more between two consecutive iterations. This affects how fast the algorithm converges, but also how accurate the bias field estimate is. As such, we see that by increasing the number of labels in the model, and if we allow the variance of these labels to update, we also make the algorithm slower because the steps taken during optimization become smaller. One way to get around this is to fix the variance to a user-defined value, which is exactly what the N3 algorithm does with its fwhm parameter: by reducing its value, "accuracy", or model certainty, is increased at the expense of speed [Sled et al. 1998].

5.4 Bias field correction of longitudinal scans

We can extend our bias field correction framework to take advantage of the information shared between images when longitudinal scans of the same subject are available. We will henceforth assume that the images at T different time points $t = 1, \ldots, T$ have been brought to a common coordinate frame, by means of a rigid groupwise registration algorithm, e.g.,[Reuter et al. 2010; Reuter et al. 2012]. We first describe the model for $T = 2$ time points, and then show how the model is easily extended to consider $T > 2$.

5.4.1 Model

For two time points, the corresponding images are denoted $\boldsymbol{y}_1 = (y_{1,1}, \ldots, y_{1,N})^T$ and $\boldsymbol{y}_2 = (y_{2,1}, \ldots, y_{2,N})^T$, where N is the number of voxels. We first assume that an (unobserved) shared image $\bar{\boldsymbol{y}} = (\bar{y}_1, \ldots, \bar{y}_N)^T$ has been generated by a

Gaussian mixture model

$$p(\bar{\boldsymbol{y}}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\bar{y}_i|\boldsymbol{\theta}), \tag{5.50}$$

$$p(\bar{y}|\boldsymbol{\theta}) = \sum_{l=1}^{L} \mathcal{N}(\bar{y}|\mu_l, \sigma_l^2)\pi_l, \tag{5.51}$$

where $\boldsymbol{\theta}_d = (\mu_1, \ldots, \mu_L, \sigma_1^2, \ldots, \sigma_L^2, \pi_1, \ldots, \pi_L)^T$. We then assume that each time point has been generated by adding zero-mean gaussian noise with variance $\bar{\sigma}^2$ to $\bar{\boldsymbol{y}}$ and then separate bias fields $\boldsymbol{b}_t = \boldsymbol{\Phi}\boldsymbol{c}_t$, i.e., for time point $t = 1$

$$p(\boldsymbol{y}_1|\bar{\boldsymbol{y}}, \boldsymbol{c}_1) = \prod_i p(y_{1,i}|\bar{y}_i, \boldsymbol{c}_1), \tag{5.52}$$

$$p(y_{1,i}|\bar{y}_i, \boldsymbol{c}_1) = \mathcal{N}(y_{1,i} - b_{1,i}|\bar{y}_i, \bar{\sigma}^2). \tag{5.53}$$

Apart from acquisition noise, the variance $\bar{\sigma}^2$ also explains potential differences between the two images due to changes in tissue over time, as well as errors in the registration.

5.4.2 Parameter Optimization

The only difference between the two time points are the Gaussian noise added to $\boldsymbol{y}$ and the independent bias fields. Therefore, the two images can be represented as a weighed average image $\boldsymbol{u}$ that contains the "common" signal and a difference image $\boldsymbol{v}$ that only contain noise and the differential bias with respect to $\boldsymbol{u}$.

The transformation from $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ to $\boldsymbol{u}$ and $\boldsymbol{v}$ can be obtained using a simple transformation matrix $\boldsymbol{T}$

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \underbrace{\begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}}_{\boldsymbol{T}} \begin{pmatrix} y_{1,i} \\ y_{2,i} \end{pmatrix}. \tag{5.54}$$

Note that we chose T to be orthonormal such that $\boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{I}$, since this simplifies the following analysis. For now, we consider a scenario where $p(\boldsymbol{\theta}) \propto 1$. The likelihood of observing $\boldsymbol{u}$ and $\boldsymbol{v}$ given the full set of parameters $\boldsymbol{\theta} = (\bar{\sigma}^2, \boldsymbol{\theta}_d, \boldsymbol{c}_1, \boldsymbol{c}_2)^T$

is

$$
\begin{aligned}
p(u_i, v_i|\boldsymbol{\theta}) &= \int_{y_i} p(u_i, v_i|\bar{y}_i, \boldsymbol{\theta}) p(\bar{y}_i|\boldsymbol{\theta}) \, \mathrm{d}\bar{y}_i \\
&= \int_{y_i} p(y_{1,i}, y_{2,i}|\bar{y}_i, \boldsymbol{\theta}) p(\bar{y}_i|\boldsymbol{\theta}) \, \mathrm{d}\bar{y}_i \\
&= \int_{y_i} \mathcal{N}(y_{1,i} - \boldsymbol{\Phi}_i \boldsymbol{c}_1|\bar{y}_i, \bar{\sigma}^2) \mathcal{N}(y_{2,i} - \boldsymbol{\Phi}_i \boldsymbol{c}_2|\bar{y}_i, \bar{\sigma}^2) p(\bar{y}_i|\boldsymbol{\theta}) \, \mathrm{d}\bar{y}_i \\
&= \int_{\bar{y}_i} \mathcal{N}(u_i - \boldsymbol{\Phi}_i \boldsymbol{c}_u|\bar{y}_i, \bar{\sigma}^2) \mathcal{N}(v_i - \boldsymbol{\Phi}_i \boldsymbol{c}_v|0, \bar{\sigma}^2) p(\bar{y}_i|\boldsymbol{\theta}) \, \mathrm{d}\bar{y}_i \\
&= \mathcal{N}(v_i - \boldsymbol{\Phi}_i \boldsymbol{c}_v|0, \bar{\sigma}^2) \sum_{l=1}^{L} \left(\int_{\bar{y}_i} \mathcal{N}(u_i - \boldsymbol{\Phi}_i \boldsymbol{c}_u|\bar{y}_i, \bar{\sigma}^2) \mathcal{N}(\bar{y}_i|\mu_l, \bar{\sigma}_l^2) \, \mathrm{d}\bar{y}_i \right) \pi_l \\
&= \underbrace{\mathcal{N}(v_i - \boldsymbol{\Phi}_i \boldsymbol{c}_v|0, \bar{\sigma}^2)}_{p(v_i|\bar{\sigma}^2, \boldsymbol{c}_v)} \underbrace{\sum_{l=1}^{L} \mathcal{N}(u_i - \boldsymbol{\Phi}_i \boldsymbol{c}_u|\mu_l, \sigma_l^2 + \bar{\sigma}^2) \pi_l}_{p(u_i|\tilde{\boldsymbol{\theta}}_d, \boldsymbol{c}_u)}, \qquad (5.55)
\end{aligned}
$$

where we have used

$$
\begin{pmatrix} y_{1,i} - \boldsymbol{\Phi}_i \boldsymbol{c}_1 - \bar{y}_i \\ y_{2,i} - \boldsymbol{\Phi}_i \boldsymbol{c}_2 - \bar{y}_i \end{pmatrix} = \boldsymbol{T}^{-1} \begin{pmatrix} u_i - \boldsymbol{\Phi}_i \boldsymbol{c}_u - \bar{y}_i \\ v_i - \boldsymbol{\Phi}_i \boldsymbol{c}_v \end{pmatrix},
$$

and $p(u_i, v_i|\bar{y}_i, \boldsymbol{\theta}) = p(y_{1,i}, y_{2,i}|\bar{y}_i, \boldsymbol{\theta})$ since $\det(\boldsymbol{T}) = 1$. Here $\tilde{\boldsymbol{\theta}}_d$ is the same as $\boldsymbol{\theta}_d$ but with re-parameterizations $\tilde{\sigma}_t^2 = \sigma_t^2 + \bar{\sigma}^2$. Finally, we can rewrite the likelihood as:

$$
\begin{aligned}
p(\boldsymbol{u}, \boldsymbol{v}|\boldsymbol{\theta}) &= \prod_i p(u_i, v_i|\boldsymbol{\theta}) \qquad (5.56) \\
&= \underbrace{\prod_i p(v_i|\bar{\sigma}^2, \boldsymbol{c}_v)}_{p(\boldsymbol{v}|\bar{\sigma}^2, \boldsymbol{c}_v)} \underbrace{\prod_i p(u_i|\tilde{\boldsymbol{\theta}}_d, \boldsymbol{c}_u)}_{p(\boldsymbol{u}|\tilde{\boldsymbol{\theta}}_d, \boldsymbol{c}_u)}. \qquad (5.57)
\end{aligned}
$$

The MAP parameters are therefore given by

$$
\begin{aligned}
\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[p(\boldsymbol{\theta}|\boldsymbol{u}, \boldsymbol{v}) \right] = p(\boldsymbol{v}|\bar{\sigma}^2, \boldsymbol{c}_v) p(\boldsymbol{u}|\tilde{\boldsymbol{\theta}}_d, \boldsymbol{c}_u) p(\boldsymbol{\theta}) \qquad (5.58) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\log p(\boldsymbol{\theta}|\boldsymbol{u}, \boldsymbol{v}) \right] = \log p(\boldsymbol{v}|\bar{\sigma}^2, \boldsymbol{c}_v) + \log p(\boldsymbol{u}|\tilde{\boldsymbol{\theta}}_d, \boldsymbol{c}_u) + \log p(\boldsymbol{\theta}), \qquad (5.59)
\end{aligned}
$$

which shows that the parameters for $\boldsymbol{u}$ and $\boldsymbol{v}$ can be estimated separately. Taking partial derivatives of the objective function and setting to zero we obtain the following parameter updates for $\boldsymbol{c}_v$ and $\bar{\sigma}^2$:

$$
\boldsymbol{c}_v = \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi} \right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{v}, \qquad (5.60)
$$

$$\bar{\sigma}^2 = \frac{\sum_{i=1}^{N}(v_i - \boldsymbol{\Phi}_i \boldsymbol{c}_v)^2}{N}, \tag{5.61}$$

which is closed form. The parameters $\boldsymbol{c}_u$ and $\tilde{\boldsymbol{\theta}}_d$ can be estimated from the weighed average image $\boldsymbol{u}$ by means of GEM as described in Section 5.2.

Once both bias field components $\boldsymbol{c}_u$ and $\boldsymbol{c}_v$ have been computed, the M bias field coefficients for the original images $c_{1,1}, \ldots, c_{1,M}, c_{2,1}, \ldots, c_{2,M}$ cain be obtained simply be a transformation back from the $\boldsymbol{u}$,$\boldsymbol{v}$ space: $(c_{1,m}, c_{2,m})^T = \boldsymbol{T}^{-1}(c_{u,m}, c_{v,m})^T$. The equations show that each image is affected by a global bias field that is common to both, and a difference field that is unique to each image.

5.4.3 Prior on the Bias Field Parameters

When we have non-uniform priors on the bias field $p(\boldsymbol{c}_1)$ and $p(\boldsymbol{c}_2)$ with the same quadratic form and same covariance matrix, we can show similar to before that they are independent with respect to $\boldsymbol{u}$ and $\boldsymbol{v}$. We incorporate a Gaussian prior on the bias field coefficients as before: $p(\boldsymbol{c}) \propto \exp(-\boldsymbol{c}^T \boldsymbol{\Psi} \boldsymbol{c})$. We further assume that the prior is applied to the coefficients of each time point independently: $p(\boldsymbol{c}_1, \boldsymbol{c}_2) = p(\boldsymbol{c}_1)p(\boldsymbol{c}_2)$. We consider an example where we have only one basis function, although it is also valid for $M > 1$ basis functions. We denote the resulting variance ψ, which yields the following joint probability

$$\begin{aligned}
p(c_u, c_v) &\propto \exp\left[-\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}^T \begin{pmatrix} \psi & 0 \\ 0 & \psi \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \right] \\
&\propto \exp\left[-\begin{pmatrix} c_u \\ c_v \end{pmatrix}^T (\boldsymbol{T}^T)^{-1} \begin{pmatrix} \psi & 0 \\ 0 & \psi \end{pmatrix} \boldsymbol{T}^{-1} \begin{pmatrix} c_u \\ c_v \end{pmatrix} \right] \\
&\propto \exp\left[-\begin{pmatrix} c_u \\ c_v \end{pmatrix}^T \begin{pmatrix} \psi & 0 \\ 0 & \psi \end{pmatrix} \begin{pmatrix} c_u \\ c_v \end{pmatrix} \right] \\
&\propto \underbrace{\exp(-c_u^2 \psi)}_{p(c_u)} \underbrace{\exp(-c_v^2 \psi)}_{p(c_v)},
\end{aligned}$$

using $\boldsymbol{T}\boldsymbol{T}^T = \boldsymbol{I}$, i.e., we penalize the bias field coefficients in the $\boldsymbol{u}, \boldsymbol{v}$ space the same way as in the original image space.

Deriving the lower bound of Equation 5.59 and setting partial derivatives to zero, we obtain the following bias field coefficient update for the difference image $\boldsymbol{v}$

$$\boldsymbol{c}_v = \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi} + \lambda 2 \bar{\sigma}^2 \boldsymbol{\Psi}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{v}, \tag{5.62}$$

We see that the updates for $\boldsymbol{c}_v$ and $\bar{\sigma^2}$ now are interdependent, which means they need to be estimated similar to GEM. This is easily done by alternating between fixing one and updating the other and vice versa, until some convergence criterion, e.g., the relative change in variance goes below some threshold. As before, the parameters for the common image $\boldsymbol{u}$ can be estimated using GEM.

5.4.4 More Than Two Time Points

The only difference between having more than two time points is that $T-1$ images $\boldsymbol{v}_t$, $t = (1, 2, \ldots, T-1)$ that form a linear combination of the original images need to be computed, whereas $\boldsymbol{u}$ remains the weighed average of all time points. The challenge is to pick the correct orthogonal $\boldsymbol{T}$ matrix that changes basis of the measurements such that we obtain these images. We can obtain this matrix using Gram-Schmidt orthogonalization.

5.5 Contributions

In this section the contributions within generative modeling of bias field correction are presented. Section 5.5.1 presents paper B, where we showed how the popular N3 algorithm can be formulated using a generative model. We then present the contributions of paper C in Section 5.5.2. Section 5.5.3 summarizes the longitudinal model for bias field correction that was previously discussed, and Section 5.5.4 discusses the software for bias field correction "IIC" that was implemented as a result of all contributions.

5.5.1 N3 Unveiled: A Heuristic MAP Estimator

In [Larsen et al. 2014] (paper B), we presented how the popular N3 bias field correction algorithm can be explained as a generative model that uses $L = 200$ Gaussians to model $\boldsymbol{u}$, but where mixture model parameters ($\boldsymbol{\pi}$) are estimated by means of a heuristic update, in particular a regularized least-squares fit.

Whereas we initially expected that N3 bias field correction would suffer due to the non-optimal parameter estimation, our experiments revealed that the mixture model fits obtained are in fact reasonable, and that bias field estimates, corrected data and corresponding mixture model fits to the histogram are comparable between N3 and EM when the hyper parameters (number of

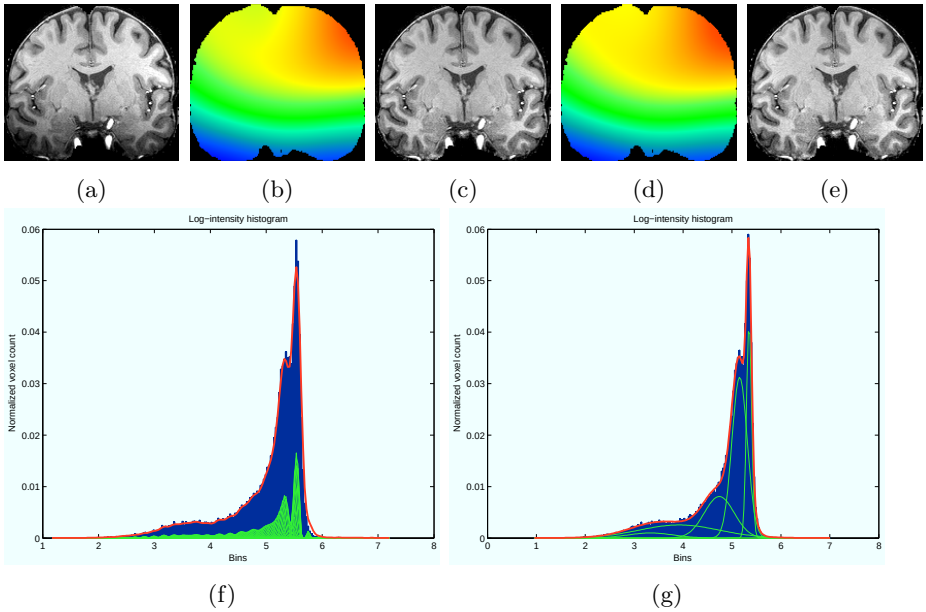


Figure 5.9: Illustrations of bias field correction of a 7T MR image: uncorrected data (a), estimated bias field and corrected data using N3 (b,c) and EM with $L = 6$ Gaussians (d,e), histogram fits at convergence using N3 (f) and EM (g) (green curves represent individual mixture components, red curve represent the full mixture model).

basis functions M, and the smoothing regularization λ) are tuned properly. This has been illustrated in Figure 5.9. Similarly, we showed that performance, measured by means of the CJV were comparable between the models for tuned hyper-parameters.

However, we observed that computational speed of the bias field correction was 3-6 times faster when we fit a mixture model composed of few Gaussians (e.g., $L = 3, 6, 9$) to the data using GEM, as compared to the $L = 200$ Gaussians fitted using the heuristic in N3 (using our own implementation in Matlab).

5.5.2 A Unified Model for Bias Field Correction

This work had a number of goals:

- To present a unified, generative model for bias field correction that is

highly configurable. We already discussed the model in Section 5.1.5. Details on how the model can be configured, and the relative merits and weaknesses of each configuration, have been elaborated in paper C.

- To introduce a version of the model that also accounts for spatial proximity when the likelihood for a voxel to belong to a label is determined. The model is inspired by SLIC superpixels [Achanta et al. 2012] and aims to remove the need for brainmasking and the use of probabilisitic atlases, both of which are particularly important at high field strengths $\geq$ 7T.
- To evaluate performance of the model configurations using a number of measures.

Given the bias field correction literature presented in chapter 4, it appears that this framework is unique, in the sense that no other works have proposed a fully generative model for bias field correction which fulfills the following conditions:

- fully utilizes GEM for the optimization of all involved parameter estimates,
- allows for, but does not require, the use of a probabilistic atlas.

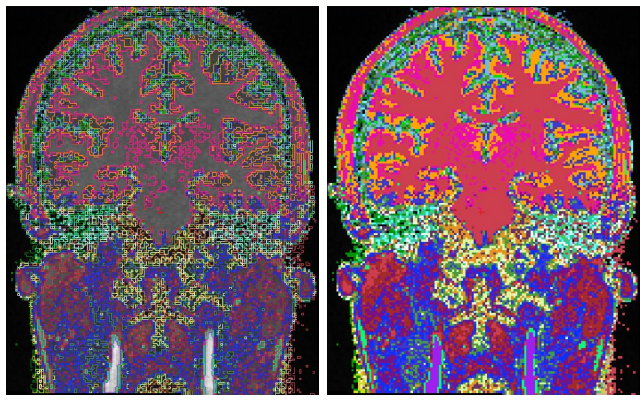
Supervoxels

Paper C presents a specific model instantiation of the generative framework which takes into account that close spatial proximity between voxels make them more likely to have been generated from the same label. The model instantiation was inspired by SLIC Superpixels [Achanta et al. 2012], and bears a lot of resemblance to the model presented in [Greenspan et al. 2006].

Specifically, we consider a multivariate Gaussian distribution where the covariance matrix is constrained to be non-zero along the diagonal only, such that each dimension is independent. We then define the likelihood of observing both the intensity and spatial position of a voxel given a label:

$$p(\boldsymbol{u}_i|l, \boldsymbol{\theta}_d) = \mathcal{N}\left(\boldsymbol{u}_i|l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\right), \tag{5.63}$$

with $\boldsymbol{\theta}_d = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_L, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_L)^T$ and exploiting notation, as $\boldsymbol{u}_i = (d_i - b_i, \boldsymbol{x}_i^T)^T$ now is a vector containing the "true" intensity as before as well as the spatial location $\boldsymbol{x}_i$. $\boldsymbol{\mu}_l$ is the mean of the multivariate distribution, and $\boldsymbol{\Sigma}_l = \text{diag}(\sigma^2_{l,intensity}, \sigma^2_{l,x}, \sigma^2_{l,y}, \sigma^2_{l,z})$ is the covariance matrix which encodes the spread of intensities and spatial positions for each supervoxel, separately for each dimension. We found that this model performed well when the spatial variance for all supervoxels was kept fixed to the initialized values.



(a) (b)

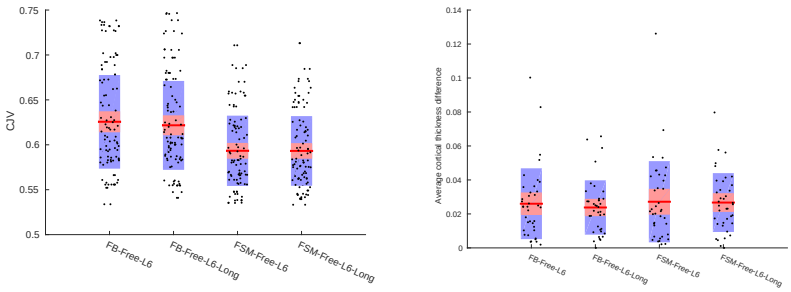
Figure 5.10: A 3T dataset segmented with the supervoxel mixture model using an initial grid spacing of 50mm.(a) outline, (b) filled segmentations of supervoxels.

Figure 5.10 illustrates a segmentation of a 3T image into the most probable supervoxels $\text{argmax}_l\, w_l^i$. In this configuration, $\sigma_\nu^2 \boldsymbol{I}$ was fixed to the squared distance between centroids at initialization (50mm^2). The supervoxel model and its parameter estimation are elaborated upon in paper C.

Evaluating Bias Field Correction Performance

Paper C contains an extensive test setting for a large number of bias field correction configurations. The model configurations were compared in terms of the CJV between gray an white matter, following leave-one-out cross-validation of the optimal regularization hyper-parameter (λ) value leading to the best mean CJV in the training set. We generally observed superior performance when employing supervoxels or a tissue atlas for correction, both at 3T and 7T.

We also measured robustness of the configurations by measuring the difference in cortical thickness given FS segmentation of a 3T scan-rescan data set composed of two coregistered time point scans per subject, taken between two days and six months apart. However, this did not translate to better robustness when measuring performance using cortical thickness in FS. Whereas some scans were acquired some months apart, the result suggests that the FS software is mostly sensitive towards configuration of the bias field (number of basis functions, regularization) as was shown in [Zheng et al. 2009], whereas the employed mixture model is less important as long as the mixture model fit is reasonable.



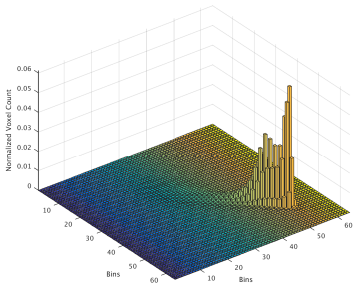
(a) (b)

Figure 5.11: Box plot showing cross-sectional vs. longitudinal model performance. (a) Measured using CJV between white and gray matter, and (b) difference in estimated cortical thickness in a 3T dataset composed of subjects with two time point scans each. Lower values equates to better performance. The red line represents the mean, while the blue box covers one standard deviation of the data and the red box covers the 95% confidence interval of the mean. The optimally corrected scan, as measured by the CJV, was fed to Freesurfer in order to obtain the cortical thickness measures. The illustrated model configurations all used $L = 6$ Gaussians, and measured were obtained after correction using (for both figures) a foreground-background mask (FB-Free-L6 and FB-Free-L6-Long) and a Freesurfer generated mask (FSM-Free-L6 and FSM-Free-L6-Long).

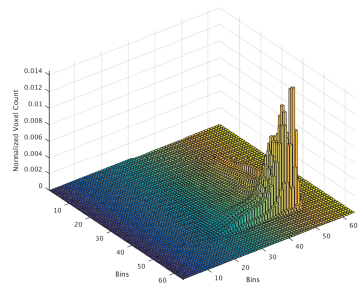
5.5.3 Longitudinal Bias Field Correction

The model for longitudinal bias field correction presented in section 5.4 is the most recent contribution in this study. Originally intended to go into paper C, it was decided to present the model here, as there is still experimental work to be done to verify the performance of the model, and further to limit the extend of paper C which was already very extensive. The model is to our knowledge the only one to correct for both the bias field that is common to all time point scans, as well as the difference bias field between scans.

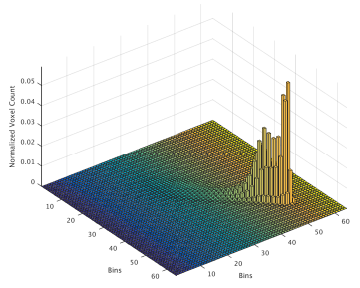
Figure 5.11 shows cross-sectional vs. longitudinal model performance on a 3T dataset composed of subjects with two time point scans each, using $L = 6$ Gaussians in the mixture model. Performance was measured using CJV between white and gray matter, and cortical thickness difference. As seen, there is no difference in performance between the cross-sectional and longitudinal model.



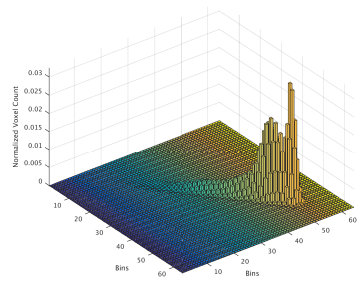
(a)



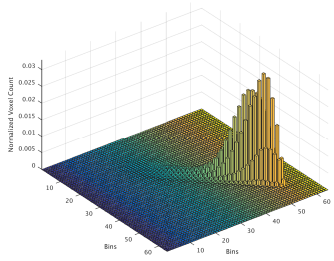
(b)



(c)



(d)



(e)

Figure 5.12: 2D histograms (normalized) of two coregistered and bias field corrected time point scans of the same subject: (a) data corrected using the cross-sectional model, (b) joint probability distribution of the data using the cross-sectional model, (c) data corrected using the longitudinal model, (d) joint probability distribution of the data given the longitudinal model, (e) uncorrected data. The longitudinal model (d) is clearly a more realistic model for the underlying data $\boldsymbol{u}$.

Figure 5.12 show 2D histograms of the uncorrected and corrected time point scans for a single subject using the cross-sectional and longitudinal correction, as well as histograms of corresponding mixture models. The figure clearly illustrate how the longitudinal model is a more realistic model, as it takes advantage of the fact that the data arises from the same underlying image (the same subject), and not from two entirely separate subjects (the assumption underlying cross-sectional correction).

5.5.4 The Software: Intensity Inhomogeneity Correction

All models presented here and in papers B and C were implemented in a Matlab software package named "Intensity Inhomogeneity Correction" (IIC). A lot of focus during this study revolved around completing the software, such that it would be a usable alternative to e.g., the N3 algorithm. To our knowledge, it is the only available software of its kind to offer bias field correction using generative models, in Matlab, and without a dependency on registration to a target template (e.g., [Ashburner and Friston 2005]). The software is complete with descriptions and help text, and is highly configurable. Paper C offers a more complete overview of the model configurations that can be used for bias field correction. Finally, a fair amount of time was spent to make the method run efficiently, emphasizing speed and limiting memory consumption wherever possible.

Efficient estimation of the bias field coefficients

The separability of the basis functions in Equation 5.21 can be exploited. It is not necessary to explicitly compute the Kronecker product $\boldsymbol{\Phi}$ listed in Equation 5.21, which is costly, both in terms of computational time and memory consumption. However, one only has to compute $\boldsymbol{\Phi}_x$, $\boldsymbol{\Phi}_y$ and $\boldsymbol{\Phi}_z$ as well as their respective Hadamard products $\boldsymbol{\Phi}_x \circ \boldsymbol{\Phi}_x$, $\boldsymbol{\Phi}_y \circ \boldsymbol{\Phi}_y$ and $\boldsymbol{\Phi}_z \circ \boldsymbol{\Phi}_z$. These are then used in a number of 1D filtering operations to obtain $\boldsymbol{\Phi}^T \boldsymbol{S} \boldsymbol{\Phi}$ and $\boldsymbol{\Phi} \boldsymbol{S} \boldsymbol{r}$ that are both necessary in order to estimate the bias field coefficients:

$$\boldsymbol{c} \leftarrow (\boldsymbol{\Phi}^\top \boldsymbol{S} \boldsymbol{\Phi} + 2\lambda \boldsymbol{\Psi})^{-1} \boldsymbol{\Phi}^\top \boldsymbol{S} \boldsymbol{r}.$$

This approach only requires a fraction of memory compared to computation of the full Kronecker product, and it is much faster. The details of this operation can be inspected in the software, which involves a lot of matrix juggling using reshape and permute.

5.5.5 Brainmasking in Freesurfer

Studies, e.g.,[Boyes et al. 2008] have shown that brainmasking is important for the quality of the bias field correction. FS uses the N3 for bias field correction, but in the current version 5.3, the brainmask *is not* used. Due to the results found in this thesis, the upcoming Freesurfer version 6.0 has the pipeline rearranged such that bias field correction is done using N3 with a brainmask and with tuned bias field hyper-parameters.

5.6 Discussion

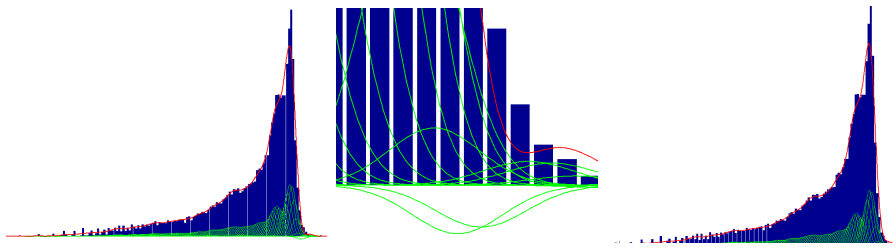
5.6.1 N3: A Box of Secrets

To verify our hypothesis about the model underlying N3, it was necessary to make our own N3 implementation in Matlab, and then compare results to the original N3 implementation. This proved challenging, in particular with respect to the cubic B-spline smoothing scheme and associated bending energy regularization, which is used to model the bias field. Furthermore, the N3 implementation is composed of several details not mentioned in the original article, which further complicated the process of verifying that it is indeed based on a generative model. In the following, we discuss some of the observations we made about the original N3 implementation throughout our analysis.

First, the regularized least-squares fit, used to fit the mixture model to the data, results in negative mixture model coefficients. This is not valid in a Gaussian mixture model, for which reason these negative coefficients are zeroed in the algorithm. This in turn results in a non-optimal fit. This was already presented in [Larsen et al. 2014], and has been illustrated in Figure 5.13.

Second, (also mentioned in the paper) after the bias field estimate is exponentiated back to the original intensity domain (following convergence of the algorithm), the estimate is smoothed one more time using the same bias field hyper-parameters that were used during bias field estimation. This is a violation of the model, as the smoothing only yields optimal bias field coefficient estimates in the log-domain of the data. The result is a bias field estimate that is no longer optimal. At least in our tests, this operation penalized performance (measured by CJV between WM and GM).

Third, N3 suffers from a minor numerical underflow in the binary that performs the regularized least-squares fit of the basis functions to the residual. This does



(a) (b) (c)

Figure 5.13: Regularized least-squares fitting of the mixture model to the histogram in N3, before zeroing of negative values of $\boldsymbol{\pi}$ (a,b), and after (c).

not violate the model, because it just results in a shift in the voxel intensities which corresponds to scaling the intensity of the bias field slightly up, but otherwise preserves its shape. This problem proved to be very subtle, and resulted in extensive debugging of our Matlab code as well as the original N3 binaries, before the problem was fully identified. In our experiments in papers B and C, we used our own implementation of the N3 smoothing scheme for all model configurations, as it does not suffer from numerical problems.

Fourth, N3 uses some fitting of Legendre polynomials to extend the bias field estimate into voxels that were masked out during the fit. We did not investigate the impact of this on performance, as we made sure that only voxels within the mask were used for our tests. In any case, voxels that have been bias field corrected using these Legendre bias field estimates are most likely not optimal.

These particularities makes it somewhat of a dilemma to use the N3 algorithm, although it has proven to work well. Following the rationale of "the proof is in the pudding", it can be argued that it is very usable for bias field correction.

5.6.2 Cross-Validating Parameters for the Bias Field

Smoothness on the bias field can be imposed both in terms of the number of basis functions M, but also by adjusting the regularization hyper-parameter λ. All tests throughout this study showed that properly tuning the bias field hyper-parameters are essential in obtaining the best possible bias field correction. This result is intriguing, because it means that one essentially need to cross-validate a two-dimensional set of parameters for every new scanner and MR-sequence that is used for image acquisition. Cross-validation of just one parameter (in particular the regularization) is very time consuming: one needs to evaluate

performance, using e.g., the CJV as a proxy[2], in a representative set of volumes from the dataset for a grid interval of parameter test values. The optimal hyper-parameters can then be determined by using e.g., a leave-one-out cross-validation strategy. This approach was employed in both papers B and C.

Testing this extensively is obviously not possible to do for every dataset available in every study, but it does suggest that more work should be spent into determining optimal hyper-parameters. A lot of studies are quite likely suffering from sub-optimal hyper-parameters and consequent impaired bias field correction.

5.6.3 Configuring the Unified Model

Whereas there are almost infinite possibilities for configuring the generative model for bias field correction, only some seem to have significant impact on its quality. In the following, we discuss those that we believe to be most relevant.

Model Flexibility

Both papers B and C showed that choosing the number of Gaussians (labels) too small (e.g., $L = 3$) penalized bias field correction performance. This is a result of too few degrees of freedom in the model, leading to a poor mixture model fit. The result is not surprising, as the data is composed of voxels not only containing WM, GM and CSF, but also partial volume effects and possibly also skull and dura, depending on the (lack of) skull-stripping. We generally observed that $L = 6$ Gaussians were sufficient to obtain a good mixture fit.

If L is chosen too high the model may become too flexible, which may result in one or several Gaussians that 'get stuck' on just a few voxels containing e.g., particular high intensities. This may lead to slower bias field correction because the variance moves towards zero, thereby reducing the size of the steps taken during each iteration of the optimization. It is further possible for a Gaussian to collapse entirely (variance becomes zero), which makes the probability for a voxel to belong to that particular label go to infinity, and in turn breaks down the optimization. This can be prevented by introducing a prior distribution on the variance parameter, thereby regularizing the values it can take. Alternatively, it easily prevented in a slightly less "elegant" way, by ensuring the variance

[2]Which in turn depends on either manual selection of WM and GM voxels, or alternatively an automated segmentation like FS.

never goes below a certain threshold; the parameter simply becomes fixed at this point.

As presented in paper C, the mixture model can be configured such that the mean parameters are optimized to be equidistantly spaced, and/or alternatively so that a single variance is used for all Gaussians. Both of these approaches help to prevent against these flexibility issues, and are coincidentally similar to how N3 handles its mixture of $L = 200$ Gaussians (fixed variance, equidistant means).

Label Prior

We observed in paper C that informing the model with a tissue atlas helped to obtain good bias field estimates and that it sped up parameter estimation considerably. The result is sensible, as the model is much more certain about the label probabilities in each voxel, and therefore makes more informed (correct) estimates.

While the performance achieved using a label prior may be negligible at 3T, we saw that it becomes much more important at high field strengths (e.g., 7T) where the bias field effect is much more prominent. Again this is sensible, as a severe bias field will make intensities of different tissues appear similar, and the prior helps the model to distinguish between them. In a worst case scenario where the label prior is not used, the gray and white matter voxels may collapse into the same label, and in this case, performance (CJV) will be worse than the original data. This may also happen if the data quality is very bad (noise in the intensities, moving artifacts, etc.).

Interestingly, model configurations similar to N3 – such as equidistantly spaced means with equal variance – may help to remedy these problems, in the sense that the bias field estimate never becomes really good (the model cannot distinguish voxels properly), but it never becomes really bad either (the voxels are never allowed to fully collapse into one label).

Supervoxels

In Figure 5.10 that illustrates an MR image segmented with supervoxels, we saw that some take up a major part of the image (e.g., white matter), whereas others cluster tightly around a few voxels composed of skull or dura. This happens because the spatial variance was kept fixed (e.g., 50mm^2) while variance of the

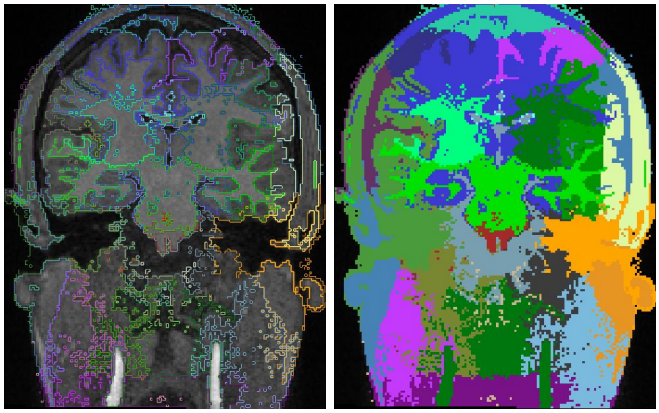


Figure 5.14: A 3T dataset segmented with the supervoxel mixture model using an initial grid spacing of 50mm using fixed mixture model coefficients.

label intensities were allowed to be updated. This makes the model emphasize voxel intensities over spatial position when the posterior probabilities for a voxel to belong to a label are computed.

A more equal spacing and sizing of supervoxels can be achieved by fixing the intensity variance and scaling it properly relative to the spatial variance (similar to what is done in [Achanta et al. 2012]), or to fix the mixture model coefficients $\pi_l = 1/L$. Figure 5.14 shows an example of supervoxels fit using the latter strategy, although preliminary testing has shown this approach to perform worse than when the spatial variance is kept fixed. Due to time constraints, we did not fully explore bias field correction using supervoxels configured this way.

In any case, the current configuration proved to work better or comparably to correction utilizing a tissue atlas. However, this performance comes at the expense of computational time, as the model suffers from very slow convergence due to the extensive amount of model parameters. The supervoxel model shows great promise, and we believe it is possible to further improve correction performance, or at the very least, computational time.

5.6.4 Longitudinal Bias Field Correction

The model covered in section 5.4 assumes a perfect registration between images. This assumption is easily violated, as no registration algorithm (to our knowledge) guarantees this. It is easy to see why the assumption is critical: if the images are not perfectly aligned, the information mutual between the images

will not add up on the same voxels, which consequently will distort bias field estimates. This assumption, and any differences between images that arise because it is not met, will be expressed in the intensity variance in the difference image, together with any differences in biology due to the distance between time points. How critical this assumption is for achieving good performance is not yet fully explored, but Figure 5.11 shows, at the very least, that performance is comparable. Further investigation is necessary, e.g., similar to [Reuter et al. 2012], with more datasets where we are certain about the time line for all scans.

Finally, it is worth mentioning that it is only necessary to fit the mixture model once to the common signal using the longitudinal model. This saves computational time, and the gain only increases as time points are added.

Chapter 6

Future Work

Potentials for future work have already been touched upon in previous chapters. Here, we summarize and elaborate upon them.

6.1 N4ITK Validation

It is a logical next step to investigate and validate the software and model underlying the N4ITK algorithm by [Tustison et al. 2010]. They present N4ITK as an evolution of N3, where the underlying cubic B-spline smoothing scheme has been adapted with a more elaborate scheme where control points are allowed to adapt to the image. However, when the generative model behind N3 is considered, the parameter estimates for the bias field coefficients already follows the optimal optimization. This means that the smoothing scheme in N4ITK replaces a valid parameter optimization with a heuristic one, *unless* the more elaborate scheme also can be explained in terms of a generative model.

[Tustison et al. 2010] suggest that N4ITK performs better than N3 given the correlation between bias fields estimated from the Brainweb image generator for a varying number of noise levels and bias field "strengths", and the ground truth. However, the results are not conclusive. First, N3 outperforms N4ITK at

the (realistic) noise level of 5% for bias fields that have been scaled in amplitude to field strengths somewhere between 1.5T and 3T.

Second, the default N3 parameters were trained on 1.5T data, *exactly* the field strength where the method outperforms N4ITK on the Brainweb data. Ideally, this training involves cross-validating the optimal distance parameter (number of cubic B-splines) and regularization hyper-parameter. As presented in [Larsen et al. 2014], these parameters, in particular the regularization, need to be re-tuned at different field strengths and scanners in order to obtain optimal performance, and N3 does not perform optimally at 3T using the default hyper-parameter value. This relationship between the number of basis functions and regularization, and its effect on bias field smoothness, is not considered by [Tustison et al. 2010]. As a result, N3 is, in our opinion, not tested in an optimal way.

Third, the smoothing schemes in the two methods are inherently different, which means you cannot compare the two using the same control point spacing hyper-parameter and expect that performance is comparable. Again, the solution is to employ a cross-validation strategy as suggested.

Finally, the bias fields generated by the Brainweb simulator are not physically correct. While the test setup with respect to the test data is the same for both methods, and therefore can be considered "fair", it remains interesting to compare the methods (including a true generative model implementation) on real MRI data, using e.g., the CJV between WM and GM as the performance measure.

6.2 Improving the Bias Field Model

Smoothing Schemes

Somewhat related to the N3 and N4ITK validation, it could be of interest to explore how well different smoothing schemes perform given proper tuning, i.e., using different basis functions and regularization covariance matrices. Theory does not suggest that one smoothing scheme is superior to another, as the only model constraint is that the field must be smooth. However, the shape of the smoothing "kernel" for a particular voxel will differ between schemes, and it is therefore possible that one scheme lends itself better to capturing subtle variations in the bias field.

Respecting the Underlying Physics

Another aspect to consider is the assumption of an entirely multiplicative field. We know this assumption to be wrong, but it is not clear exactly how much it affects correction performance, in particular at high field strengths $\geq$ 7T. Therefore, it could be of interest to integrate a more physically correct model in a generative framework.

6.3 Extending the Supervoxel Model

The supervoxel model was derived relatively late in the course of this PhD study, which leaves a lot of potential for further research. In particular, the model is highly configurable, and it has not been uncovered how this should be done in order to achieve optimal performance. The following configuration aspects should be considered in this regard:

- Initialization (supervoxel mean spacing, variance and weight).
- The relationship between intensity and spatial variance (should one or both be kept fixed as is implicitly done in [Achanta et al. 2012], should one be scaled with respect to the other, etc.).
- Determining if mixture coefficients should be the same and fixed throughout the optimization process.

Furthermore, the presented model is just a first step in incorporating spatial proximity between voxels and labels in the model, without depending on Markov random fields or anatomical atlases. Even though the model already appears to be very powerful, it may be possible to make the model even more elaborate, thereby achieving even better bias field correction.

6.4 Computational Speed

Resource consumption is always important. Our current implementation "IIC" has already been optimized in some aspects, in particular with respect to the residual smoothing (bias field coefficient estimation). However, the code can most definitely be improved to consume less memory and run faster. This is

particularly important when the number of labels is high, as is the case for the supervoxel model which is currently very slow (computational time and number of iterations to achieve convergence summarized in paper C).

As previously discussed, the nature of EM results in optimization steps that always guarantee a better solution than before. For the purpose of resource consumption, the order and arrangement of the two different components of the M-step in GEM optimization are important. While it seems reasonable to ensure that the mixture model has been properly fitted to the data before the bias field is estimated, there may be certain arrangements that yields particular good computational performance without impacting correction quality. For example, it may be enough to fit the mixture model using a fixed number of iterations, rather than allowing it to fully converge given e.g., the relative or absolute change in cost per iteration. Another example is model configurations where the variance is the same for all labels. In these configurations, the smoothing scheme can be further optimized for an increase in computational speed.

Finally, the current implementation is in Matlab, which is suboptimal. For the method to be truly useful to research, it has to be implemented in C++ for improved computational speed and reduced memory consumption. The nature of this work is mostly practical, but proper code optimization is not trivial.

6.5 Longitudinal Bias Field Correction

The longitudinal model currently performs comparably to it's cross-sectional counterpart on one dataset in terms of correction quality. Whereas the model should always guarantee a decrease in computational time which is linear with respect to the number of time points corrected, it is yet to be fully determined how well the model performs, in particular in datasets where time point scans for each subject are both close and far apart in time, in datasets containing pathology and in datasets of different field strengths.

The model currently assumes registration prior to correction, which potentially limits performance and usability. It would therefore be interesting to integrate registration in a more unified model, which has already been explored to some extent, i.e., by [Ashburner and Ridgway 2013].

6.6 Integration in Freesurfer

It is the intention that IIC will make its way into the FS pipeline and replace the N3 algorithm. This depends in part on obtaining a successful (re)-implementation of the software in C++, and also on extensive testing.

Chapter 7

Conclusion

In this thesis, we presented aspects of both development and application of tools for MRI analysis, which have lead to a number of contributions within the field of MRI analysis of the brain.

First, it was presented how the software Freesurfer was used to successfully analyze a dataset obtained in the study ADEX, exploring the effects of moderate-to-high aerobic exercise in patients with mild-to-moderate Alzheimer's disease. While it was not possible to show that four months of exercise leads to a significant reduction in brain atrophy or improved cognition, findings did indicate that exercise correlates with volumetric brain changes, and that changes in frontal cortical thickness correlate with changes in cognitive performance, measured using the Symbol Digits Modalities and Verbal Fluency Tests. This work has been presented in paper A.

Second, it was described how the Freesurfer software relies on the popular bias field correction algorithm N3. We presented how it is important that the software is tuned properly with respect to a number of hyper-parameters, in order to obtain optimal bias field correction given a number of measures, and how this may affect study outcomes such as those presented in the ADEX study. We further showed in paper B how the N3 algorithm can be fully explained within a generative modeling framework, with specific parameters being updated using heuristic estimation techniques. We used this, together with an overview of

current literature on bias field correction to motivate research into generative models for bias field correction.

Third, the research into generative models for bias field correction has resulted in a fully developed method for bias field correction in Matlab named "Intensity Inhomogeneity Correction", which is freely available, and which can be run without supplying other input than the MRI data to be corrected. Furthermore, we present the generative framework underlying the correction method which is, by itself, a contribution. In the framework we included a new generative model that encodes spatial proximity between image voxels and label centers using a Gaussian probability distribution, thereby enabling correction of data that typically requires an anatomical atlas at high field strengths $\geq$ 7T. This work is presented in depth in the paper C manuscript.

Finally, we presented and discussed a model for bias field correction of longitudinal time point scans of the same subject, correcting both the bias that is common to all scans, and also the bias from the difference image. It is the intention that this model will be the focus in another journal paper.

In the ideal world, the presented method for bias field correction would have been implemented in C++ and integrated in the Freesurfer pipeline, thereby combining application and development of MRI analysis tools and emphasizing how the two areas depend on each other, as was discussed in the introduction. However, it was not possible to make this achievement within the timespan of the PhD study. This, together with a number of proposed suggestions for further research into bias field correction, are therefore left as future work.

Paper A

Effect of moderate-to-high intensity aerobic exercise on hippocampus and cortical regions in patients with mild to moderate Alzheimer's disease

C. T. Larsen[12], K. S. Frederiksen[3], S. G. Hasselbalch[3], A. N. Christensen[2], P. Høgh[4], L. Wermuth[5], A. Lolk[5], B. B. Andersen[3], H.R. Siebner[1,6], G. Waldemar[3], E. Garde[1].

[1]DANISH RESEARCH CENTRE FOR MAGNETIC RESONANCE, COPENHAGEN UNIVERSITY HOSPITAL HVIDOVRE, DENMARK

[2] DEPARTMENT OF APPLIED MATHEMATICS AND COMPUTER SCIENCE, TECHNICAL UNIVERSITY OF DENMARK, KONGENS LYNGBY, DENMARK

[3]DANISH DEMENTIA RESEARCH CENTER, DEPT. OF NEUROLOGY, RIGSHOSPITALET, UNIVERSITY OF COPENHAGEN, DENMARK

[4]REGIONAL DEMENTIA RESEARCH CENTER, REGION ZEALAND, ROSKILDE HOSPITAL, UNIVERSITY OF COPENHAGEN, DENMARK

[5]DEMENTIA CLINIC, ODENSE UNIVERSITY HOSPITAL, ODENSE, DENMARK

[6] DEPARTMENT OF NEUROLOGY, COPENHAGEN UNIVERSITY HOSPITAL BISPEBJERG, COPENHAGEN, DENMARK

Corresponding author

Christian T. Larsen

Danish Research Center for Magnetic Resonance

Section 714

Copenhagen University Hospital

Hvidovre

Kettegaard Allé 30

2650 Hvidovre

Denmark

cthla@dtu.dk

Abstract

Background: Studies on healthy elderly have shown that aerobic exercise has a positive effect on both brain structure and function. So far studies in patients with Alzheimer's disease (AD) are few and results have been inconsistent. In this study, we wanted to assess the relationship between aerobic exercise, brain changes measured by MRI and cognitive functioning in patients with AD.

Methods: As part of a larger randomized controlled trial this MR-sub-study included forty-two patients. For both control and exercise group MR and cognitive assessment was performed at baseline and after 16 weeks with 60-minutes exercise sessions three times a week. Both attendance and intensity were monitored providing a total exercise load. Changes in regional brain volumes and cortical thickness were analysed using Freesurfer and volume of white matter hyperintensities (WMH) quantified.

Results: Exercise load showed a positive correlation with changes in volume in the hippocampal subfields, as well as frontal, cortical thickness in the exercise group. Changes in frontal, cortical thickness correlated with measures of mental speed and attention (SDMT) and verbal fluency in both groups. Volume of WMH were associates with changes in hippocampal volume.

Conclusion: In patients with AD the effect of exercise on hippocampal volume appear to depend on training attendance and intensity. The extent of WMH may modify the effect of physical training but further studies are needed.

Introduction

Alzheimer's disease (AD) is a neuro-degenerative disease, characterized by progressive impairment of memory [GM84] and atrophy of specific brain regions, in particular the hippocampus [JB07]. Atrophy of the anterior hippocampus can be observed in patients with mild cognitive impairment (MCI) as early as three years prior to onset of AD, with increasing involvement of the hippocampus as the symptoms progresses [JW07].

In addition, atrophy has been observed in the amygdala, entorhinal cortex and fusiform gyrus in MCI, progressing to the middle temporal gyrus, posterior temporal lope and parietal lobe in patients with AD [JW07].

The effect of current pharmacological treatments of AD are at best symptomatic [RC12] but recent studies suggest that non-pharmacological approaches such as physical exercise may have a beneficial effect on cognitive functioning as well as brain structure.

In healthy elderly both cognition, physical functioning and performance in activities of daily living were improved when given a home training program consisting of daily exercises and walking [AV12]. Studies including MRI suggest that physical training is associated with increased whole brain volume [SC06], less atrophy in frontal, parietal and temporal cortex [SC03] and even an increase in grey matter volume in pre-frontal and cingulate cortex [RR11]. In preadolescent children Chaddock et al [LC10] showed a relation between basal ganglia volume, enhanced cognitive functioning and aerobic fitness while in healthy elderly improvement in memory function were associated with increase hippocampal volume [KE11].

In patients with AD Andrade et al reports an increase in frontal cognitive function, after following a multimodal exercise program for 16 weeks [LA13].

Increasing evidence support that exercise benefits brain function and structure but also that there are multiple pathways and that age and concurrent pathological processes may modify the effect. In a recently published study exercise had a positive effect on neuropsychiatric symptoms and cognition in a relatively large group of patients with mild to moderate (KH 2016). To our knowledge, this is the first to investigate the effect of continuously supervised moderate to high-intensity exercise program in patients with mild to moderate AD. In a sub-study, MRI was performed at baseline and 16-week follow-up in order to assess the effect on regional brain volumes. The primary outcome measure is hippocampal volume and we hypothesize that in AD patients a moderate to intense exercise program will preserve hippocampal volume.

Methods

Participants and study design

The primary objective of the ADEX study was to assess the effect of moderate-to-high intensity aerobic exercise on cognitive and physical functioning, quality of life and ADL in two-hundred community-dwelling patients with mild to moderate AD. The participants were randomized into a control and exercise group, the latter performing 60 minutes of moderate-to-high-intensity aerobic exercise three times weekly for sixteen weeks. Psychological, cognitive and physical performance was assessed before and after the sixteen week period for both groups. The trial ran for 5 rounds from 2012 to 2014.

The procedure for screening, as well as inclusion and exclusion criteria, is described in a recent publication [KH13]. In brief, key inclusion criteria included age between 50 and 90 years and a Minimal Mental State Examination (MMSE) score of more than 19, whereas exclusion criteria included presence of medical and psychiatric diseases, alcohol abuse and regular, weekly high-intensity exercise.

A subgroup consisting of seventy-one patients from memory clinics in Copenhagen, Roskilde and Odense was invited for brain MRI at baseline and 16 week follow-up. Thirteen of these patients left the study prematurely, and sixteen patients were excluded due to poor MRI data quality (movement artefacts) (9), data processing problems (6) or notable error in data processing outcomes (1), leaving 42 patients for the present study.

The ADEX trial was approved by the The Committees of Biomedical Research Ethics for the Capital Region (Protokol no.: H-3-2011-128) and by the Danish Data Protection Agency (j.no.: 30-0718).

MRI acquisition

Both baseline and follow-up MRI was performed at Hvidovre Hospital, Denmark, using a 3.0-T Siemens Tim Trio scanner and included T1-weighed magnetization-prepared rapid gradient echo (MPRAGE) (TE 3.04ms, TR 1550ms, FoV read 256mm, FoV phase 100%, 192 slices), T2-weighed fast spin echo (TE 354ms, TR 3000ms, FoV read 282mm, FoV phase 76.6%, 192 slices) and fluid attenuated inversion recovery (FLAIR) (TE 353ms, TR 6000ms, FoV read 282mm, FoV phase 85.9%, 192 slices) sequences.

Data processing

Regional, cortical thickness and hippocampal volume

The T1-weighed data was gradient unwarped to correct for spatial distortions [JC06], and then processed with version 5.3 of the cross-sectional [BF02] and longitudinal [MR12] Freesurfer stream, in order to obtain segmentations of cortical regions defined according to the Destrieux atlas [CD10] as well as the hippocampal subfields [KL09], caudate and putamen. The pipeline was specifically tuned to correct for intensity inhomogeneity that can be observed at 3T [RB08, WZ09].

In cases where Freesurfer failed to properly delineate the white matter and pial surface, the pipeline were manually guided following the steps outline in the Freesurfer documentation (http://freesurfer.net/fswiki/FreeSurferWiki). This specifically involved correcting the skull stripping to better delineate the pial surface, insertion of control points to guide white matter normalization for the purpose of improving white matter segmentation, and finally editing the white matter segmentation itself. Two trained readers edited the pipeline; to avoid segmentation bias, one was

responsible for skull stripping and white matter editing, while the other was responsible for control point insertion.

Finally, overall quality of the longitudinal segmentation output were asserted by experienced raters (CTL, KSF, EG). Specifically, the pial and white matter surface outlines, as well as the hippocampal subcortical segmentation were visually inspected and consensus reached for all. One volume was excluded due to significant segmentation error in the hippocampus.

To explore regional, cortical effects, gyri and sulci thickness measures obtained from Freesurfer were divided into four categories (early, middle, late, and very late) each including areas reported to be progressively affected by atrophy from mild cognitive impairment MCI to full AD diagnosis [JW09]: 'early' (temporal, precuneus, cingulate), 'middle' (parietal, temporal-occipital, occipital, fusiform, parahippocampus) and 'late' (frontal). A 'very late' region composed by the pre and postcentral cortex were also defined (supplementary material, table 4).

Whole and parenchymal brain volume

Freesurfer also provides measures of brain volume (BV), brain parenchymal volume (BPV), white matter volume (WM) and intracranial volume (ICV). Whole brain volume included all segmented structures, excluding background and the brain stem. Parenchymal volume further excluded the ventricles (lateral, inferior lateral, 3rd, 4th and 5th), CSF and choroid plexus. Brain parenchymal fraction (BPF) was obtained by dividing BPV with ICV.

White matter hyperintensities

For delineation of white matter hyperintensities (WMH), MPRAGE and T2-weighted images were co-registered and re-sliced to the corresponding FLAIR image using a 6 parameter rigid transformation. WMH were defined as clearly hyperintense areas relative to surrounding white matter on both FLAIR and T2-weighted images and identified by simultaneous inspection of both aligned images. For WMH volume local thresholding was applied and WMH volumes for the whole brain quantified automatically using the Jim image analysis package, Version 6.0, (Xinapse Systems Ltd., Northants, UK, www.xinapse.com). Visual identification and delineation was carried out by a single trained rater blinded to clinical information. For nine subjects (five control, four intervention) WMH could not delineated due to movement artefacts.

Longitudinal and normalized measures

Longitudinal measures of brain volume, cortical thickness and cognitive scores for each subject were computed as the *relative* change between baseline and follow-up by subtracting baseline from follow-up, and dividing the difference with the baseline measure, thereby canceling out within-subject correlations, as well as accounting for between-subject differences in brain size. Throughout the paper, we will refer to the relative change simply as *change*.

A normalized WMH measure was obtained by dividing WMH volume with white matter volume.

Cognitive outcome measures

Cognitive assessment included the Minimal Mental State Examination (MMSE) for global cognitive impairment [MF75], the Symbol Digit Modalities Test (SDMT) for mental speed and attention [AS82] (only measurements at 120 seconds included in the analysis), and the Stroop Color and Word Test (Stroop) incongruent score for reaction time. Verbal memory performance was assessed by the Alzheimer's Disease Assessment Scale – Cognitive Subscale (ADAS-Cog) [WR84], and verbal fluency (VFT) as number of words produced over 1 minute each ([KH15] for details).

Exercise load (attendance and intensity)

To assess training attendance and intensity a training log was created. Attendance was logged, and attendance ratio defined as number of attended exercise sessions over total number of offered sessions. Exercise intensity was based on the per-session average heart rate (HR) recorded using continued monitoring during exercise (including rest). Average HR for all sessions was calculated, and intensity defined as average HR over maximum expected HR (220 minus subject age). To obtain total exercise load, measures for attendance ratio and intensity was multiplied.

Statistical analysis

Brain volume measures

Separate multivariate models were used to compare changes in volume between groups for the hippocampal subfields (model 1), para-hippocampus (model 2), caudate and putamen (model 3). Similarly, separate models were used to compare changes in thickness of the cortical gyri and sulci respectively for each of the 'early', 'middle', 'late' and 'very late' categories previously described.

For all group tests, Hotellings T^2 multivariate test [HH31] was applied. Hotellings T^2 eliminates the need for testing each individual measure in a model (e.g., each of the hippocampal subfields), and consequently the need for performing a multiple comparisons test. This makes the test more sensitive and less prone to type II error (false negatives) than e.g., univariate tests with Bonferoni correction.

Since outliers were detected in scatter plots of the variables, a further non-parametric Oja rank test [HO04] were performed, using 10.000 permutations, to confirm validity of p-values from Hotellings T^2 test.

Correlation tests

The relationship between changes in hippocampal subfield volume and verbal memory (ADAS-Cog measure) was tested with caudate and putamen as a control regions. Also, the relationship between changes in frontal and cingulate cortical thickness, and verbal memory, mental speed and attention (SDMT, VFT, Stroop) were assessed with pre- and postcentral cortex as a control region. The relationship between cognitive measures and changes in volume of caudate and putamen were explored with hippocampus as a control volume. Finally, relationship between exercise load and changes in hippocampal subfield volume as well as frontal cortical thickness were also investigated for the exercise group only. (Details on cortical labels can be seen in the supplementary material, table 5).

All correlation tests were performed by computing the covariance matrix between changes in test scores and brain measures, and then testing the nul-hypothesis of zero covariance between the two. This yields an overall p-value for the full covariance matrix, but not an r-value. If the overall p-value was significant, a post-hoc analysis of the correlation for each single brain measure was performed, yielding individual r- and p-values.

Six subjects were excluded from the correlation tests due to missing cognitive scores (one subject lADAS-Cog; three subjects SDMT; one subject VFT; six subjects Stroop).

For all tests, the significance level was 0.05. Gender, age and baseline WMH were used as covariates. Statistics were obtained with SAS Statistical Software version 9.4 and Rstudio 2.15.2

Results

No significant differences were found for any baseline characteristics between the control and intervention group (Table 1).

Brain volumes

In the hippocampal subfield model a significant difference were found for the left fimbria (p=0.012) and CA2_3 (p=0.016) which however, could not be found when correcting for multiple comparisons (figure 1, table2). No difference between groups was observed for the parahippocampal or caudate and putamen models (table 2). No significant between-group difference in changes in regional cortical thickness was found (table 3).

The normalized WMH measure did not change significantly from baseline to follow-up (p=0.996). In both groups WMH was associated with changes in the hippocampal subfield volume (p=0.002), specifically presubiculum (r=-0.345, p=0.031), and CA4_DG (r=-0.406, p=0.010), (supplement, table 6) as well as changes in both gyri (p=0.048) and sulci (p=0.002) thickness in the 'very late' category, and also sulci thickness in the 'middle' category (p=0.0495). Inspection of the individual significant gyri and sulci showed the largest correlations with the right post-central gyri (r=0.221, p=0.170) in the ‘very late’ gyri, the right post-central sulci (r=0.467, p=0.002) in the ‘very late’ sulci category, and the right occipital superior and transversal sulci (r=0.460, p=0.003) in the ‘middle’ category.

Cognitive performance correlations

Change in the frontal and cingulate cortical thickness correlated significantly with both SDMT (p=0.025) and VFT (p=0.026), (table 7, supplement). Specifically, for SDMT a moderate correlation was found with change in cortical thickness of the right frontal inferior-orbital gyri (r=0.464, p=0.004) and right frontal inferior-triangular gyri (r=0.386, p=0.020). Per-group investigation (figure 2) revealed moderate correlations in both regions for the exercise group but not the control group. For VFT, correlations was found with cortical thickness changes in the left (r=0.384, p=0.017) and right (r=0.328, p=0.044) frontal mid-posterior gyri and sulci. Per-group investigation of this relationship (figure 3) revealed significant correlations in both regions for the control group but not the exercise group. (table 8, supplement).

A separate analysis of the correlation between normalized WMH and each cognitive performance measure showed a significant correlation with Stroop (r=0.394, p=0.023). Inspection of the scatterplot of the measures revealed the correlation to be dominated by two points with no apparent overall trend. No significant correlations were found between WMH and the SDMT/VFT measures.

Exercise load correlations

Exercise load was found to associate significantly with changes in cortical thickness in the frontal cortex (p=0.0106), especially for the right frontal inferior sulci (r=0.514, p=0.034). Similarly, a significant correlation with changes in volume in the hippocampal subfields (p=0.0091) was found, the strongest correlation showing in the right subiculum (r=0.443, p=0.086).

Discussion

To our knowledge, this is the first study to explore the effects of supervised moderate-to-high intensity aerobic exercise on regional brain atrophy measures in patients with mild to moderate AD.

Main findings in this study are three-fold. First, exercise load shows a positive correlation with changes in volume in the hippocampal subfields, as well as frontal, cortical thickness which support that exercise does stimulate brain growth, which agrees with previous findings, e.g., [SC03], [SC06] and [KE11]. The group differences in changes in the left fimbria and CA2_3 hippocampal subfield volume disagree (figure 1). While changes in the left fimbria could suggest that exercise stimulates brain growth with measurable effects already after 16 weeks in patients with AD, the opposite trend in CA2_3 suggest that the effects are spurious.

We performed a post-hoc qualitative inspection of changes in hippocampal, caudate and putamen volume and regional cortical thickness. Our observations suggested only a minor loss of tissue across subjects across a four month period, and furthermore that loss of tissue seems to be less in exercise participants. In some cases, the data suggested a slight increase. However, given the variation in data and the inability to show significant differences between groups, this remains only partially indicative of the effects of exercise.

Previous literature, e.g., Erickson et al, who showed an effect in global hippocampal volume in healthy elderly after 1-2 years of PE [KE11]. Our findings suggests that 16 weeks may not be enough to effect duration as well as intensity should be considered when planning exercise programs. An alternative explanation for the disagreement is that the control group may have been exercising outside of the study, thereby diminishing group differences.

Second, for all participants, changes in frontal, cortical thickness were associated with SDMT and VFT. Cortical thinning have previously been shown to associate with cognitive impairment, e.g., in Parkinson's Disease [BS14], which suggests that a decline in cortical thickness can be used as an indicator of progressive cognitive impairment. Our finding that frontal, cortical thickness associates with SDMT/VFT cognitive performance measures, is in agreement with this, and more generally with literature on the role the frontal cortex has in mental speed, attention and verbal fluency [JA06].

Given that AD is a neuro-degenerative disease, it would be expected that changes in cortical thickness and cognitive measures would be primarily negative. However, no particular decline or increase in cortical thickness or cognitive performance were observed for changes in frontal, cortical thickness and the SDMT/VFT scores (figure 2 and 3). Further inspection reveals a stronger relationship between changes in cortical thickness and SDMT for the exercise group, while the control group exhibits the strongest correlation with VFT.

Third, interestingly, normalized WMH values were associated with changes in hippocampal volume as well as regional, cortical thickness in both groups, suggesting that pathology other then AD may influence brain structure. WMH are generally regarded a marker of small vessel disease and severe WMH an indicator of poor vascular health. Recent findings (presented at AAIC 2015) suggest that the presence of WMH may modify the effect of exercise intervention but also that exercise may enhance vascular health as well as connectivity.

A recent study [EL15] have shown that the rate of change in WMH are strongest during conversion from MCI to AD diagnosis, and follows the rate of change in hippocampal volume suggesting that WMH may play a part in the conversion process. Similarly, Caso et al [CF15] conclude that WM degeneration may be an early marker of pathological changes in atypical AD. In line with these studies, our finding that the severity of WMH associates with relative changes in the hippocampal subfields implies a relationship between WMH and the rate of neuro-degeneration in AD.

WMH has also been shown to associate with cognitive decline [OC10] and more specifically medial temporal atrophy, attention and frontal executive functions [YS11] as well as frontal atrophy and

reduced delayed recall performance [IM12]. Here, we observed positive correlations between WMH and changes in cortical thickness, which is not what we would expect. The finding was made predominantly in the cortical control regions, where effects due to AD would be expected to show very late. As such, the finding could be spurious, and may again suggest that a four month period is not be enough to measure significant differences with controls.

Although the underlying mechanism is still not completely understood animal studies indicate that physical exercise stimulates neurogenesis and formation of brain-derived neuro-protective factor (BDNF) [HP05][KN09][MM13]. This relationship between exercise and BDNF has also been reported in humans [CC02].

Methodological considerations

The disagreement between findings in the left fimbria and left CA2_3 hippocampal subfields, the positive correlations between WMH and regional, cortical thickness suggest some level of uncertainty in the statistics, one explanation being noise. Furthermore, the population size after drop-outs, data processing and quality assurance in this study was quite small, which affects statistical power and consequently increases the likelihood of spurious results.

From a clinical perspective, sixteen weeks of exercise are likely not enough to measure effects on the same level as e.g., Erickson et al [KE11]. Furthermore, previous findings have been in healthy elderly subjects, with a presumably small pathological load. It is very likely the effects of an exercise intervention on the brain might be different in patients with AD.

The subjects in this study were recruited from three different memory clinics. This could pose a potential bias in statistics, given that subjects may be treated differently at each center. We did not control for this, due to the already mentioned statistical concerns. Furthermore, there is a good agreement between the statistics of the control and intervention group as shown in Table 1. Given that subjects from all centers using a controlled, randomized process, it seems reasonable to assume that differences between subjects due to memory centers even out.

As described in [MR12], the longitudinal Freesurfer pipeline utilizes cross-sectionally processed time points to generate a common template, which is then used as a point of initialization for an unbiased analysis of each individual timepoint. This procedure helps to avoid potential bias in the outcome measures due to e.g., registration to the baseline time point, as pointed out in [NF11]. Furthermore, it increases statistical power because inter-subject variation is reduced.

Conclusion

A sixteen week intervention of moderate-to-high aerobic exercise clearly translated to observable changes in hippocampal subfield volume and cortical thickness in a group of patients with AD. The finding supports evidence that exercise stimulates exercise, which may potentially have a negating effect on neuro-degeneration.

Furthermore, correlations between changes in the frontal cortex and mental speed and verbal fluency show that changes in brain volume and cortical thickness does relate to changes in cognition. As such, it seems likely that exercise for prolonged periods of time may increase the extent to which brain growth is stimulated, thereby leading to a positive effect on cognition and ADL in patients with AD.

Finally, an observed association between the extent of WMH and changes in the hippocampus suggest that WMH may be indicative of the rate of neuro-degeneration. It is further suggested that this may have a limiting effect on the effectiveness of exercise.

Future studies should explore the effect of aerobic exercise focusing on a prolonged duration of the intervention period, as well as an increase in the number of participants receiving MRI, in order to increase statistical power.

Acknowledgements

The ADEX study is supported by a grant from The Danish Council for Strategic Research (j. no.: 10-092814).

Investigator list by site:

Memory Clinic, Copenhagen University Hospital, Rigshospitalet

Birgitte Bo Andersen, DMSc, M.D.

Memory Clinic, Roskilde Hospital

Peter Høgh, PhD, M.D.

Dementia Clinic, Odense University Hospital

Anette Lolk, Assoc. Professor, PhD, M.D.

Lene Wermuth, M.D.

Department of Geriatrics, Svendborg Hospital

Søren Jakobsen, M.D.

Department of Geriatrics, Slagelse Hospital

Lars P. Laugesen, M.D.

Robert Graff Gergelyffy, M.D.

Memory Clinic, Aarhus University Hospital

Hans Brændgaard, M.D.

Hanne Gottrup, PhD, M.D.

Memory Clinic, Aalborg Hospital

Karsten Vestergaard, M.D.

Memory Clinic, Glostrup University Hospital

Eva Bjerregaard, M.D.

We are grateful to all physiotherapists, study nurses and clinical raters for their contributions to the study. We also thank Jonathan Polimeni from the Athinoula Martinos Center, Massachussetts General Hospital, Boston, USA for supplying gradient unwarping software.

Figures and tables

Table 1: Baseline demographics for all participants and basic volumetric measures for participants in the MR sub-study.

	MR-substudy			Main study	
	Control (N=20)	**Exercise (N=22)**	**p-values**	**Control (N=93)**	**Exercise (N=107)**
Age (years), *mean (±SD)*	69 (7.5)	68 (7.7)	0.681	71.3 (7.3)	69.8 (7.4)
Gender, *male/female (N)*	12 / 8	14 / 8	0.809	57/36	56/51
MMSE, *median (±SD)*	26.0 (2.3)	26.0 (3.4)	0.712	24.1 (3.8)	23.8 (3.4)
Hypertension*, *N (%)*	5 (25)	4 (18)	0.591	35 (37.6)	48 (44.9)
WMH ($mm^3 \cdot 10^3$), *median (±SD)*	0.83 (6.0)	0.86 (6.1)	0.183	N/A	N/A
WMH/WM (10^{-3}), *median (±SD)*	6.00 (15.8)	2.80 (12.2)	0.085	N/A	N/A
BV ($mm^3 \cdot 10^6$), *mean (±SD)*	1.01 (0.13)	1.11 (0.12)	0.489	N/A	N/A
BPV ($mm^3 \cdot 10^6$), *mean (±SD)*	1.00 (0.11)	1.05 (0.11)	0.226	N/A	N/A
BPF, *mean (±SD)*	0.63 (0.03)	0.64 (0.03)	0.456	N/A	N/A

* Hypertension was defined as ≥140/90 mmHg.

WMH: White matter hyperintensities, WM: white matter, BV: brain volume, BPV:brain parenchymal volume, BPF: brain parenchymal fraction.

Table 2: Difference between control and exercise group in change in volume measures.

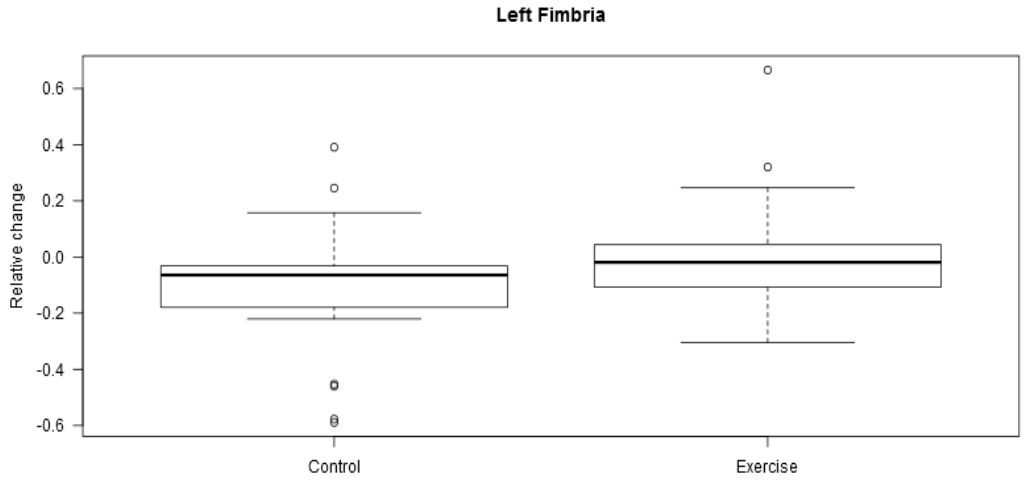
		Left side (p-value)	Right side (p-value)
Test 1	Hippocampus		
	Total hippocampus volume *	0.930	0.266
	Presubiculum	0.747	0.817
	Subiculum	0.485	0.469
	Fimbria	0.012	0.423
	Hippocampal fissure	0.240	0.583
	CA1	0.470	0.237
	CA2_3	0.016	0.299
	CA4_DG	0.120	0.184
Test 2			
	Parahippocampus	0.462	0.962
Test 3			
	Caudate	0.128	0.071
	Putamen	0.289	0.779

* Includes the 'hippocampus' class, which captures voxels that were not put in the other subfield categories.

Table 3: Difference between control and exercise group in change in cortical, regional thickness.

	Gyri (p-value)	Sulci (p-value)
Early	0.255	0.187
Middle	0.112	0.623
Late	0.687	0.126
Very late	0.121	0.506

Figure 1: Boxplots showing the changes in volume in the left fimbria and left CA2_3 hippocampal subfields for the control and exercise groups.



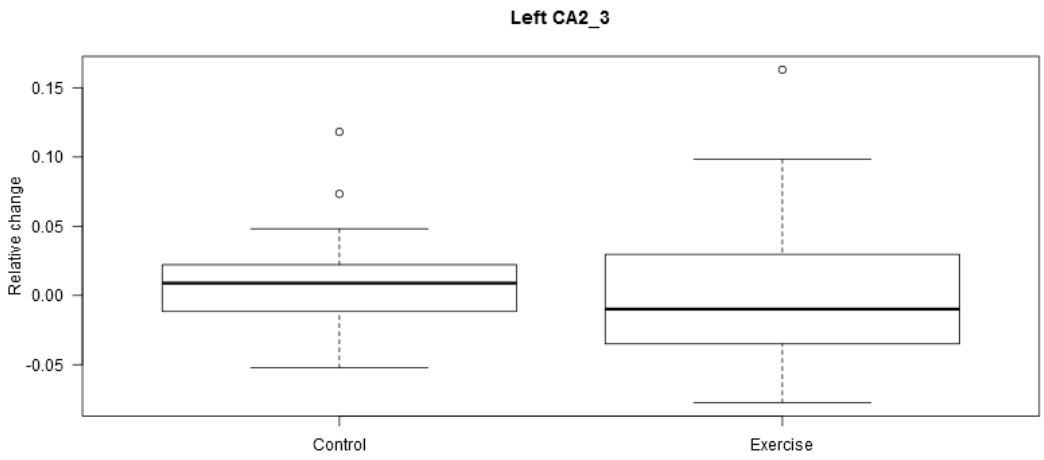


Figure 2: Correlation between change in the right frontal inferior-orbital/inferior-triangular gyri thickness and relative change in SDMT measured at 120 seconds.

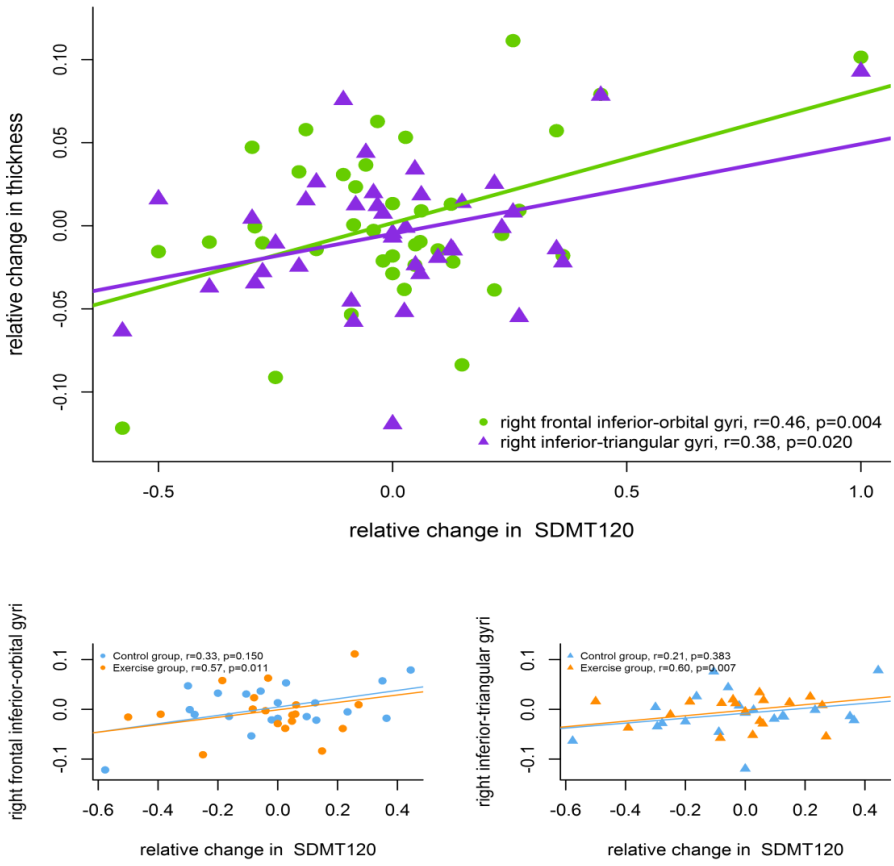
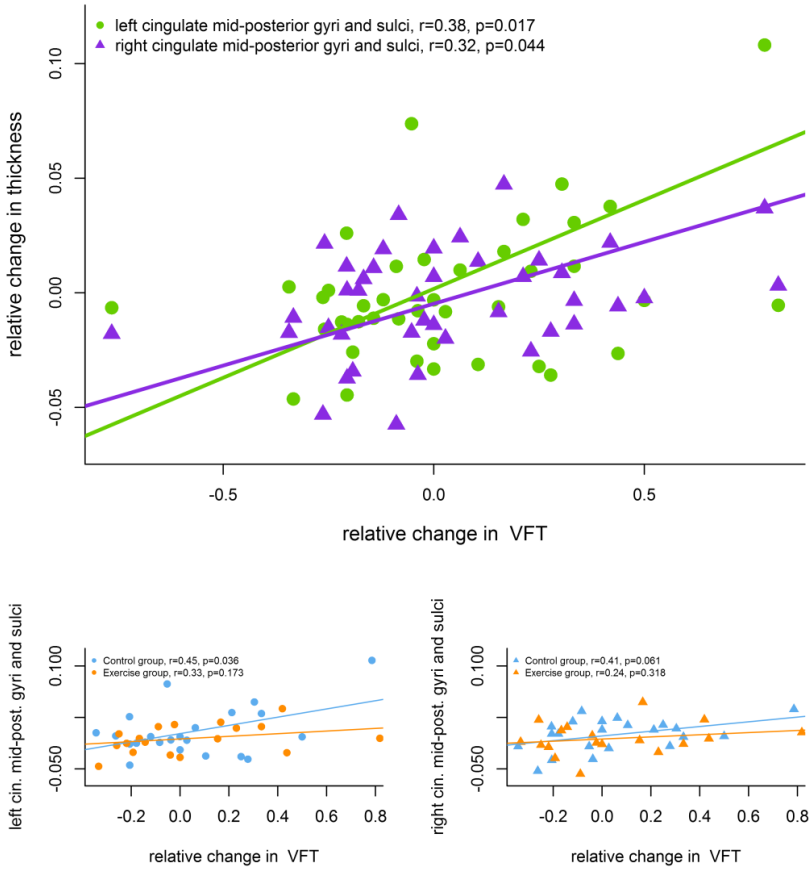


Figure 3: Correlation between relative change in the left and right cingulate, mid-posterior gyri and sulci thickness and relative change in VFT.



References

[JA06] Alvarez JA, Emory E: Executive function and the frontal lobes: A meta-analytic review. Neuropsychology Review 2006;16(1);17–42.

[LA13] Andrade LP de, Gobbi LTB, Coelho FGM, Christofoletti G, Costa JLR, Stella F: Benefits of Multimodal Exercise Intervention for Postural Control and Frontal Cognitive Functions in Individuals with Alzheimer's Disease: A Controlled Trial. J Am Geriatr Soc. 2013;61(11):1919-26.

[JB07] Barnes J, Godbolt AK, Frost C, Boyes RG, Jones BF, Scahill RI, Rossor MN, Fox NC: Atrophy rates of the cingulate gyrus and hippocampus in AD and FTLD. Neurobiol Aging 2007;28(1):20-28.

[CC02] Cotman CW, Berchtold NC. Exercise: a behavioral intervention to enhance brain health and plasticity. *Trends Neurosci*. 2002 Jun;25(6):295-301 .

[RC12] Castellani RJ, Perry G: Pathogenesis and disease-modifying therapy in Alzheimer's disease: the flat line of progress. Arch Med Res 2012;43: 694–698.

[FC15] Caso F, Agosta F, Mattavelli D, Migliaccio R, Canu E, Magnani G, Marcone A, Copetti M, Falautano M, Comi G, Falini A, Filippi M: White Matter Degeneration in Atypical Alzheimer Disease. Neuroradiology 2015.

[LC10] Chaddock L, Erickson KI, Prakash RS, VanPatter M, Voss MW, Pontifex MB, Raine LB, Hillman CH, Kramer AF: Basal Ganglia Volume is associated with Aerobic Fitness in Preadolescent Children. Developmental Neuroscience 2010;32;249-256.

[OC10] Carmichael O, Schwarz C, Drucker D, Fletcher E, Harvey D, Beckett L, Jack CR Jr, Weiner M, DeCarli C: Longitudinal changes in white matter disease and cognition in the first year of the Alzheimer disease. Arch Neurol. 2010;Nov;67(11):1370-1378.

[SC03] Colcombe SJ, Erickson KI, Raz N, Webb AG, Cohen NJ, McAuley E, Kramer AF: Aerobic Fitness Reduces Brain Tissue Loss in Aging Humans. J Gerontol A Biol Sci Med Sci. 2003;58(2):176-80.

[SC06] Colcombe SJ, Erickson KI, Scalf PE, Kim JS, Prakash R, McAuley E, et al: Aerobic exercise training increases brain volume in aging humans. J Gerontol A Biol Sci Med Sci 2006;61:1166–1170.

[JC06] Jovicich J, Czanner S, Greve D, Haley E, Kouwe A, Gollub R, Kennedy D, Schmitt F, Brown G, MacFall J, Fischl B, Dale A: Reliability in Multi-Site Structural MRI Studies: Effects of Gradient Non-linearity Correction on Phantom and Human Data. NeuroImage 2006;30(2):436-43.

[CD10] Destrieux C, Fischl B, Dale A, Halgren E: Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. NeuroImage 2010;53(1);1-15.

[KE11] Erickson KI, Voss MW, Prakash RS, Basak C, Szabo A, Chaddock L, et al: Exercise training increases size of hippocampus and improves memory. Proc Natl Acad Sci USA 2011;108:3017–3022.

[BF02] Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, Montillo A, Makris N, Rosen B, Dale AM: Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 2002;33:341-355.

[NF11] Fox NC, Ridgway GR, Schott JM: Algorithms, atrophy and Alzheimer's disease: Cautionary tales for clinical trials. NeuroImage 2011;51:15-18.

[MF75] Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12(3);189-198.

[CG78] Golden C. Stroop color an word test manual. Chicago; Stoelting Co; 1978.

[HH31] Hotelling H. The Generalization of Student's Ratio. Ann. Math. Statist. 2 1931;3;360-378.

[KH13] Hoffmann K, Frederiksen KS, Sobol NA, Beyer N, Vogel A, Simonsen AH, Johannsen P, Lolk A, Terkelsen O, Cotman CW, Hasselbalch SG, Waldemar G: Preserving Cognition, Quality of Life, Physical

Health and Functional Ability in Alzheimer's Disease: The Effect of Physical Exercise (ADEX Trial): Rationale and Design. Neuroepidemiology 2013; 31:198-207.

[KH15] Hoffmann K, Sobol N, Beyer N, Frederiksen KS, Vogel A, Vestergaard K, Brændgaard H, Gottrup H, Lolk A, Jakobsen S, Laugesen L, Gergelyffly R, Hoegh P, Bjerregaard E, Siersma V, Andersen B, Johannsen P, Cotman C, Waldemar G, Hasselbalch S: Moderate to High Intensity Physical Exercise in Patients with Alzheimer's Disease. (submitted).

[RJ07] Johnson RA, and Wichern DW: Applied multivariate statistical analysis 6th Edition. Englewood Cliffs, NJ: Prentice hall, 2007.

[EL15] Lindemer ER, Salat DH, Smith EE, Nguyen K, Fischl B, Greve DN: White matter signal abnormality quality differentiates mild cognitive impairment that converts to Alzheimer's disease from nonconverters. Neurobiology of Aging 2015;xxx;1-15.

[KL09] Leemput KV, Bakkour A, Benner T, Wiggins G, Wald LL, Augustinack J, Dickerson BC, Golland P, Fischl B:Automated Segmentation of Hippocampal Subfields From Ultra-High Resolution In Vivo MRI. Hippocampus 2009;19;549-557.

[GM84] McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM: Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 1984;34:939–944.

[HM47] Mann HB and Whitney DR: 'On a test of whether one of two random variables is stochastically larger than the other', Annals of Mathematical Statistics 1947;18;50-60.

[IM12] Meier IB, Manly JJ, Provenzano FA, Louie KS, Wasserman BT, Griffith EY, Hector JT, Allocco E, Brickman AM: White matter predictors of cognitive functioning in older adults. J Int Neuropsychol Soc. 2012;May;18(3):414-427.

[JM99] Mu JS, Li WP, Yao ZB, Zhou XF: Deprivation of endogenous brain-derived neurotrophic factor results in impairment of spatial learning and memory in adult rats. Brain Res 1999;835:259–265.

[KM95] Korte M, Carroll P, Wolf E, Brem G, Thoenen H, Bonhoeffer T: Hippocampal long-term potentiation is impaired in mice lacking brain-derived neurotrophic factor. Proc Natl Acad Sci USA 1995;92: 8856–8860.

[MM13] Marlatt MW, Potter MC, Bayer TA, van Praag H, Lucassen PJ: Prolonged Running, not fluoxetine Treatment, Increases Neurogenesis, but does not alter neuropathology, in the 3xTg Mouse Model of Alzheimer's Disease. Curr Top Behav Neurosci. 2013;15:313-40.

[KN09] Nichol K, Deeny SP, Seif J, Camaclang K, Cotman CW: Exercise improves cognition and hippocampal plasticity in APOE ε4 mice. Alzheimers Dement 2009; 5:287–294.

[HO04] Oja, Hannu; Randles, Ronald H. Multivariate Nonparametric Tests. Statist. Sci. 19 2004;4;598-605.

[HP05] van Praag H, Shubert T, Zhao C, Gage FH: Exercise enhances learning and hippocampal neurogenesis in aged mice. J Neurosci 2005;25:8680–8685.

[MR12] Reuter M, Schmansky NJ, Rosas, HD, Fischl B: Within-Subject Template Estimation for Unbiased Longitudinal Image Analysis. Neuroimage 2012;61(4):1402-1418.

[WR84] Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's Disease. Am J Psychiatry 1984;141(11); 1356-1364.

[RR11] Ruscheweyh R, Willemer C, Kruger K, Duning T, Warnecke T, Sommer J, Volker K, Ho HV, Mooren FG, Knecht S, Floel F: Physical activity and memory functions: An interventional study. Neurobiol Aging. 2011;32(7):1304-1319.

[AS82] Smith A. Symbol Digits Modailities Test (SDMT) Manual (revised). Los Angeles: Western Psychological Services; 1982.

[BS14] Segura B, Baggio HC, Marti MJ, Valldeoriola F, Compta Y, Garcia-Diaz AI, Vendrell P, Bargallo N, Tolosa E, Junque C.: Cortical thinning associated with mild cognitive impairment in Parkinson's disease. Mov Disord. 2014 Oct;29(12):1495-503

[YS11] Shim YS, Youn YC, Na DL, Kim SY, Cheong HK, Moon SY, Park KW, Ku BD, Lee JY, Jeong JH, Kang H, Kim EJ, Lee JS, Go SM, Kim SH, Cha KR, Seo SW: Effects of medial temporal atrophy and white matter hyperintensities on the cognitive functions in patients with Alzheimer's disease. Eur Neurol. 2011;66(2):75-82.

[AV12] Vreugdenhil A, Cannell J, Davies A, Razay G: A community-based exercise programme to improve functional ability in people with Alzheimer's disease: a randomized controlled trial. Scand J Caring Sci. 2012;26(1):12-19.

[JW07] Whitwell J, Przybelski S, Weigand SD, Knopman DS, Boeve BF, Petersen RC, Jack CR: 3D Maps from Multiple MRI Illustrate Changing Atrophy Patterns as Subjects Progress from MCI to AD. BRAIN. 2007;130(7):1777-1786.

Supplementary

Cortical regions (Freesurfer, Destrieux atlas nomenclature)

Table 4: Gyri and sulci according to known progression of AD partitioned into categories of 'early', 'middle', 'late' and ‘very late’.

Gyri	
Early	temporal (inferior, medial, superior lateral), temporal (transversal, plan-polar, plan-tempo), precuneus, cingulate (transversal+ventral)
Middle	parietal (inferior-angular+supramar, superior), occipital-temporal medial (parahippocampal, fusiform)
Late	frontal (inferior-opercular/orbital/triangular, medial, superior)
Very Late	precentral, postcentral
Sulci	
Early	temporal (inferior, superior, transverse), cingulate-marginalis
Middle	parieto-occipital, occipital-temporal lateral, occipital (anterior, middle-lunatus, superior-tranversal)
Late	frontal (inferior, middle, superior)
Very Late	precentral (inferior, superior), postcentral

Table 5: Gyri and sulci of the frontal and cingulate cortex, as well as the precentral and postcentral cortex.

Effect	frontal (inferior, middle, superior) gyri and sulci, cingulate (mid-posterior, mid-anterior, anterior) gyri and sulci
Control	precentral gyri, postcentral gyri, postcentral sulci, precentral (inferior, superior) sulci

Results, additional tables

Table 6: Correlations between baseline normalized WMH and change in volume in the hippocampal subfields, parahippocampus, caudate and putamen for all study participants.

		Left side		Right side	
		r-value	p-value	r-value	p-value
Test 1	***Hippocampus***				
	Sum of all subfields	-0.263	0.106	0.147	0.373
	Presubiculum	**-0.345**	**0.031**	0.238	0.145
	Subiculum	-0.295	0.068	0.263	0.106
	Fimbria	0.281	0.083	0.009	0.956
	Hippocampal fissure	-0.225	0.169	-0.313	0.053
	CA1	-0.098	0.552	0.058	0.726
	CA2_3	0.042	0.802	0.085	0.607
	CA4_DG	**-0.406**	**0.010**	0.132	0.423
Test 2					
	Parahippocampus	-0.058	0.721	0.135	0.407
Test 3					
	Caudate	0.301	0.059	0.204	0.208
	Putamen	-0.140	0.390	0.012	0.941

Table 7: correlations between changes in caudate, putamen volume/frontal, cingulate cortical thickness and changes in SDMT/VFT/Stroop outcome measures for all participants.

		Effect (p-value)	Control (p-value)
Frontal, cingulate cortex			
	SDMT	**0.0246**	0.6749
	Stroop	0.8873	0.5966
	VTF	**0.0259**	0.4788
Caudate, Putamen			
	SDMT	0.7042	0.1978
	Stroop	0.8424	0.7530
	VTF	0.0801	0.3846

Table 8: Correlations between relative changes in frontal/cingulate cortical thickness and VFT/SDMT outcome measures for all participants.

	VFT (r-value)	VFT (p-value)	SDMT (r-value)	SDMT (p-value)
Left:				
Frontal inferior sulci	0.037	0.825	0.103	0.549
Frontal middle sulci	0.182	0.275	0.187	0.275
Frontal superior sulci	0.249	0.132	0.099	0.566
Frontal inferior-opercular gyri	0.297	0.071	-0.069	0.690
Frontal inferior-orbital gyri	0.164	0.326	-0.057	0.743
Frontal inferior-triangular gyri	0.319	0.051	-0.032	0.854
Cingulate mid-posterior gyri and sulci	**0.384**	**0.017**	0.118	0.494
Right:				
Frontal inferior sulci	0.174	0.297	0.167	0.324
Frontal middle sulci	0.276	0.093	-0.179	0.296
Frontal superior sulci	0.227	0.170	0.208	0.224
Frontal inferior-opercular gyri	0.269	0.103	0.279	0.104
Frontal inferior-orbital gyri	0.114	0.494	**0.464**	**0.004**
Frontal inferior-triangular gyri	0.277	0.092	**0.386**	**0.020**
Cingulate mid-posterior gyri and sulci	**0.328**	**0.044**	-0.126	0.463

Significant p-values have been highlighted in orange, bold and the corresponding correlations in blue, bold. P-values for the individual gyri and sulci regions have not been corrected for multiple comparisons.

Paper B

N3 Bias Field Correction Explained as a Bayesian Modeling Method

Christian Thode Larsen[1], J. Eugenio Iglesias[23], and Koen Van Leemput[124]

[1] Department of Applied Mathematics and Computer Science, Technical University of Denmark
[2] Martinos Center for Biomedical Imaging, MGH, Harvard Medical School, USA
[3] Basque Center on Cognition, Brain and Language, Spain
[4] Departments of Information and Computer Science and of Biomedical Engineering and Computational Science, Aalto University, Finland

Abstract. Although N3 is perhaps the most widely used method for MRI bias field correction, its underlying mechanism is in fact not well understood. Specifically, the method relies on a relatively heuristic recipe of alternating iterative steps that does not optimize any particular objective function. In this paper we explain the successful bias field correction properties of N3 by showing that it implicitly uses the same generative models and computational strategies as expectation maximization (EM) based bias field correction methods. We demonstrate experimentally that purely EM-based methods are capable of producing bias field correction results comparable to those of N3 in less computation time.

1 Introduction

Due to its superior image contrast in soft tissue without involving ionizing radiation, magnetic resonance imaging (MRI) is the *de facto* modality in brain studies, and it is widely used to examine other anatomical regions as well. MRI suffers from an imaging artifact commonly referred to as "intensity inhomogeneity" or "bias field", which appears as low-frequency multiplicative noise in the images. This artifact is present at all magnetic field strengths, but is more prominent at the higher fields that see increasing use (e.g., 3T or 7T data). Since intensity inhomogeneity negatively impacts any computerized analysis of the MRI data, its correction is often one of the first steps in MRI analysis pipelines.

A number of works have proposed bias field correction methods that are integrated into tissue classification algorithms, typically within the domain of brain MRI analysis [1–7]. These methods often rely on generative probabilistic models, and combine Gaussian mixtures to model the image intensities with a spatially smooth, multiplicative model of the bias field artifact. Cast as a Bayesian inference problem, fitting these models to the MRI data employs expectation-maximization (EM) [8] optimizers to estimate some [7] or all [1, 3, 4, 6] of the model parameters. Specifically tailored for brain MRI analysis applications, these methods encode strong prior knowledge about the number and spatial distribution of tissue types present in the images. As such, they cannot be used out of the box to bias field correct imaging data from arbitrary anatomical regions.

In contrast, the popular N3 [9] bias field correction algorithm does not require any prior information about the MRI input. This allows N3 to correct images of various locations and contrasts, and even automatically handle images that contain pathology. However, despite excellent performance and widespread use, its underlying bias field correction mechanism is not well understood. Specifically, the original paper [9] presents N3 as a relatively heuristic recipe for increasing the "frequency content" of the histogram of an image, by performing specific iterative steps without optimization of any particular objective function.

This paper aims to demonstrate how N3 is in fact intimately linked to EM-based bias field correction methods. In particular, N3 uses the same generative models and bias field estimation computations; however, instead of using dedicated Gaussian mixture models that encode specific prior anatomical knowledge, N3 uses generic models with a very large number of components (200) that are fitted to the histogram by a regularized least-squares method.

The contribution of this paper is twofold. First, to the best of our knowledge, this is the first study offering theoretical insight into why the seemingly heuristic N3 iterations yield such successful bias field estimations. Second, we demonstrate experimentally on datasets of 3T and 7T brain scans that standard EM-based methods, using far less components, are able to produce comparable bias field estimation performance at reduced computational cost.

2 Methods

In this section, we first describe the N3 bias field correction method and its practical implementation. We then present EM-based bias field correction and the generative model it is based upon. Finally, we build an analogy between the two methods, thereby pointing out their close similarities.

2.1 The N3 method in its practical implementation

The following description is based on version 1.12[1] of the N3 method. In order to facilitate relating the method to a generative model in subsequent sections, we deviate from the notational conventions used in the original paper [9]. Furthermore, whereas the original paper only provides a high-level description of the algorithm (including integrals in the continuous domain, etc.), here we describe the actual implementation in which various discretization, interpolation, and other processing steps are performed.

Let $\boldsymbol{d} = (d_1, \ldots, d_N)^T$ be the intensities of the N voxels of a MRI scan, and let $\boldsymbol{b} = (b_1, \ldots, b_N)^T$ be the corresponding gains due to the bias field. As commonly done in the bias field correction literature [1, 3, 4, 6], N3 assumes that $\boldsymbol{d}$ and $\boldsymbol{b}$ have been log-transformed, such that the effect of $\boldsymbol{b}$ is additive. The central idea behind N3 is that the histogram of $\boldsymbol{d}$ is a blurred version of the histogram of the true, underlying image due to convolution with the histogram of $\boldsymbol{b}$, under the

[1] Source code freely available from `http://packages.bic.mni.mcgill.ca/tgz/`.

assumption that $\boldsymbol{b}$ has the shape of a zero-mean Gaussian with known variance. The algorithm aims to reverse this by means of Wiener deconvolution and to estimate a smooth bias field model accordingly. This reversal process is repeated iteratively, because it was found to improve the bias field estimates [9].

Deconvolution step: The first step of the algorithm is to deconvolve the histogram. Given the current bias field estimate denoted $\tilde{\boldsymbol{b}}$, a normalized histogram with $K = 200$ bins of bias field corrected data $\boldsymbol{d} - \tilde{\boldsymbol{b}}$ is computed[2]. The bin centers are given by

$$\tilde{\mu}_1 = \min(\boldsymbol{d} - \tilde{\boldsymbol{b}}), \quad \tilde{\mu}_K = \max(\boldsymbol{d} - \tilde{\boldsymbol{b}}), \quad \tilde{\mu}_k = \tilde{\mu}_1 + (k-1)h, \tag{1}$$

where $h = (\tilde{\mu}_K - \tilde{\mu}_1)/(K-1)$ is the bin width, and the histogram entries $\{v_k, k = 1, \ldots, K\}$ are filled using the following interpolation model:

$$v_k = \frac{1}{N}\sum_{i=1}^{N} \varphi\left[\frac{d_i - \tilde{b}_i - \tilde{\mu}_k}{h}\right], \qquad \varphi[s] = \begin{cases} 1 - |s| & \text{if } |s| < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Defining $\hat{\boldsymbol{v}}$ as a padded, 512-dimensional vector such that $\hat{\boldsymbol{v}} = (\boldsymbol{0}_{156}^T, \boldsymbol{v}^T, \boldsymbol{0}_{156}^T)^T$, where $\boldsymbol{v} = (v_1, \ldots, v_K)^T$ and $\boldsymbol{0}_{156}$ is an all-zero 156-dimensional vector, the histogram is deconvolved by

$$\hat{\boldsymbol{\pi}} \leftarrow \boldsymbol{F}^{-1}\boldsymbol{D}\boldsymbol{F}\hat{\boldsymbol{v}}. \tag{2}$$

Here $\boldsymbol{F}$ denotes the 512×512 Discrete Fourier Transform matrix with elements

$$F_{n,k} = e^{-2\pi j(k-1)(n-1)/512}, \quad n, k = 1, \ldots, 512$$

and $\boldsymbol{D}$ is a 512×512 diagonal matrix with elements

$$D_k = \frac{f_k^*}{|f_k|^2 + \gamma}, \quad k = 1, \ldots, 512$$

where γ is a constant value set to $\gamma = 0.1$, and $\boldsymbol{f} = (f1, \ldots, f_{512})^T = \boldsymbol{F}\boldsymbol{g}$. Here $\boldsymbol{g}$ denotes a 512-dimensional vector that contains a wrapped Gaussian kernel with variance

$$\tilde{\sigma}^2 = \frac{f^2}{8\log 2}, \tag{3}$$

such that

$$\boldsymbol{g} = (g_1, \ldots, g_{512})^T, \quad g_l = \begin{cases} h\mathcal{N}((l-1)h|0, \tilde{\sigma}^2) & \text{if } l = 1, \ldots, 256 \\ g_{512-l+1}, & \text{otherwise,} \end{cases} \tag{4}$$

where f denotes a user-specified full-width-at-half-maximum parameter (0.15 by default), and $\mathcal{N}(\cdot|\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2.

After $\hat{\boldsymbol{\pi}}$ has been computed by means of Eq. (2), any negative weights are set to zero, and the padding is removed in order to obtain the central deconvolved 200-entry histogram $\tilde{\boldsymbol{\pi}}$.

[2] A flat bias field: $\tilde{\boldsymbol{b}} = \boldsymbol{0}$ is assumed in the first iteration.

Bias correction step: When the histogram $\tilde{\boldsymbol{\pi}}$ has been deconvolved, the corresponding "corrected" intensity $\tilde{d}_{\mu_l}$ in the deconvolved histogram is estimated at each bin center $\tilde{\mu}_l, l = 1, \ldots, K$ by

$$\tilde{d}_{\mu_l} = \sum_k w_k^l \tilde{\mu}_k \quad \text{with} \quad w_k^l = \frac{\mathcal{N}\left(\tilde{\mu}_l | \tilde{\mu}_k, \tilde{\sigma}_k^2\right) \tilde{\pi}_k}{\sum_{k'} \mathcal{N}\left(\tilde{\mu}_l | \tilde{\mu}_{k'}, \tilde{\sigma}_{k'}^2\right) \tilde{\pi}_{k'}},$$

and a "corrected" intensity $\tilde{d}_i$ is found in every voxel by linear interpolation:

$$\tilde{d}_i = \sum_{l=1}^{K} \tilde{d}_{\mu_l} \varphi \left[\frac{d_i - \tilde{b}_i - \tilde{\mu}_l}{h} \right], \qquad \varphi[s] = \begin{cases} 1 - |s| & \text{if } |s| < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Finally, a residual $\boldsymbol{r} = \boldsymbol{d} - \tilde{\boldsymbol{d}}$ is computed and smoothed in order to obtain a bias field estimate:

$$\tilde{\boldsymbol{b}} = \boldsymbol{\Phi} \tilde{\boldsymbol{c}} \tag{5}$$

where

$$\tilde{\boldsymbol{c}} \leftarrow \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi} + N \beta \boldsymbol{\Psi} \right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{r}. \tag{6}$$

Here $\boldsymbol{\Phi}$ is a $N \times M$ matrix of M spatially smooth basis functions, where element $\Phi_{i,m}$ evaluates the m-th basis function in voxel i; $\boldsymbol{\Psi}$ is a positive semi-definite matrix that penalizes curvature of the bias field; and β is a user-determined regularization constant (the default is $\beta = 10^{-7}$).

Post-processing: N3 alternates between the deconvolution step and the bias field correction step until the standard deviation of the difference in bias estimates between two iterations drops below a certain threshold (default: $\varsigma = 10^{-3}$). By default, N3 operates on a subsampled volume (factor 4). After convergence, the bias field estimate is exponentiated back into the original intensity domain, where it is subsequently fitted with Eq. (6), i.e., with $\boldsymbol{r} = \exp(\tilde{\boldsymbol{b}})$. The resulting coefficients are then used to compute a final bias field estimate by evaluation of Eq. (5) with $\boldsymbol{\Phi}$ at full image resolution. The uncorrected data is finally divided by the bias field estimate in order to obtain the corrected volume.

2.2 EM-based bias field estimation

In the following we describe the generative model and parameter optimization strategy underlying EM-based bias field correction methods[3].

[3] Several well-known variants only estimate a subset of the parameters considered here – e.g., in [1] the mixture model parameters are assumed to be known, while [3] uses fixed, spatially varying prior probabilities of tissue types.

Generative model: Maintaining the notation $\boldsymbol{d}$ to denote a log-transformed image and $\boldsymbol{b} = \boldsymbol{\Phi c}$ to denote a parametric bias field model with parameters $\boldsymbol{c}$, the "true", underlying image $\boldsymbol{d} - \boldsymbol{b}$ is assumed to be a set of N independent samples from a Gaussian mixture model with K components – each with its own mean μ_k, variance σ_k^2, and relative frequency π_k (where $\pi_k \geq 0, \forall k$ and $\sum_k \pi_k = 1$). Given the model parameters $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_k, \sigma_1^2, \ldots, \sigma_K^2, \pi_1, \ldots, \pi_K, c_1, \ldots, c_M)^T$, the probability of an image is therefore

$$p(\boldsymbol{d}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \left[\sum_{k=1}^{K} \mathcal{N}(d_i - \sum_{m=1}^{M} c_m \Phi_{i,m} | \mu_k, \sigma_k^2) \pi_k \right]. \tag{7}$$

The generative model is completed by a prior distribution on its parameters, which is typically of the form

$$p(\boldsymbol{\theta}) \propto \exp[-\lambda \boldsymbol{c}^T \boldsymbol{\Psi} \boldsymbol{c}],$$

where λ is a user-specified regularization hyperparameter and $\boldsymbol{\Psi}$ is a positive semi-definite regularization matrix. This model encompasses approaches where bias field smoothness is imposed either solely through the choice of basis functions (i.e., $\lambda = 0$, as in [3]), or through regularization only (i.e., $\boldsymbol{\Phi} = \boldsymbol{I}$, as in [1]). The prior is uniform with respect to the mixture model parameters.

Parameter optimization: According to Bayes's rule, the maximum a posteriori (MAP) parameters are given by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p(\boldsymbol{\theta}|\boldsymbol{d}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\log p(\boldsymbol{d}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \right]. \tag{8}$$

By exploiting the specific structure of $p(\boldsymbol{d}|\boldsymbol{\theta})$ given by Eq. (7), this optimization can be performed conveniently using a generalized EM (GEM) algorithm [8, 3]. In particular, GEM iteratively builds a lower bound $\varphi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ of the objective function that touches it at the current estimate $\tilde{\boldsymbol{\theta}}$ of the model parameters (E step), and subsequently improves $\varphi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ with respect to the parameters (M step) [8, 10]. This procedure automatically guarantees to increase the value of the objective function at each iteration. Constructing the lower bound involves computing soft assignments of each voxel i to each class k:

$$w_k^i = \frac{\mathcal{N}\left(d_i - \sum_m \tilde{c}_m \Phi_{i,m} | \tilde{\mu}_k, \tilde{\sigma}_k^2\right) \tilde{\pi}_k}{\sum_{k'} \mathcal{N}\left(d_i - \sum_m \tilde{c}_m \Phi_{i,m} | \tilde{\mu}_{k'}, \tilde{\sigma}_{k'}^2\right) \tilde{\pi}_{k'}}, \tag{9}$$

which yields the following lower bound:

$$\varphi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}) = \sum_i \left[\sum_k w_k^i \log \left(\frac{\mathcal{N}(d_i - \sum_m c_m \Phi_{i,m} | \mu_k, \sigma_k^2) \pi_k}{w_k^i} \right) \right] - \lambda \boldsymbol{c}^T \boldsymbol{\Psi} \boldsymbol{c}. \tag{10}$$

Optimizing Eq. (10) simultaneously for the Gaussian mixture model parameters and bias field parameters is difficult. However, optimization with respect to the

mixture model parameters for a given set of bias field parameters is closed form:

$$\tilde{\mu}_k \leftarrow \frac{\sum_i w_k^i (d_i - \sum_m \tilde{c}_m \boldsymbol{\Phi}_{i,m})}{\sum_i w_k^i}, \quad \tilde{\sigma}_k^2 \leftarrow \frac{\sum_i w_k^i \left(d_i - \sum_m \tilde{c}_m \boldsymbol{\Phi}_{i,m} - \tilde{\mu}_k\right)^2}{\sum_i w_k^i} \tag{11}$$

$$\tilde{\pi}_k \leftarrow \frac{\sum_i w_k^i}{N}. \tag{12}$$

Similarly, for a given set of mixture model parameters the optimal bias field parameters are given by

$$\tilde{\boldsymbol{c}} \leftarrow \left(\boldsymbol{\Phi}^T \boldsymbol{S} \boldsymbol{\Phi} + 2\lambda \boldsymbol{\Psi}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{S} \boldsymbol{r}, \tag{13}$$

with

$$s_k^i = \frac{w_k^i}{\tilde{\sigma}_k^2}, \quad s_i = \sum_k s_k^i, \quad \boldsymbol{S} = \mathrm{diag}(s_i), \quad \tilde{d}_i = \frac{\sum_k s_k^i \tilde{\mu}_k}{\sum_k s_k^i}, \quad \boldsymbol{r} = \boldsymbol{d} - \tilde{\boldsymbol{d}}.$$

Valid GEM algorithms solving Eq. (8) are now obtained by alternately updating the voxels' class assignments (Eq. (9)), the mixture model parameters (Eqns. (11) and (12)), and the bias field parameters (Eq. (13)), in any order or arrangement.

2.3 N3 as an approximate MAP parameter estimator

Having laid out the details of both N3 and EM-based bias field correction, we are in a position to illustrate parallels between these two methods. In particular, as we describe below, *N3 implicitly uses the same generative model as EM methods* and shares the exact same bias field parameter update (up to numerical discretization aspects). The only difference is that, whereas EM methods fit their Gaussian mixture models by maximum likelihood estimation, N3 does so by regularized least-squares fitting of the mixture model to the histogram entries. Thus, whereas N3 was conceived as iteratively deconvolving Gaussian bias field histograms from the data without optimizing any particular objective function, its successful performance can be readily understood from a standard Bayesian modeling perspective.

Considering the generative model described in Section 2.2, we postulate that N3 uses $K = 200$ Gaussian distributions that are equidistantly spaced between the minimum and maximum intensity, i.e., the parameters $\{\mu_k\}$ are fixed (Eq. (1)). Furthermore, all Gaussians are forced to have an identical variance that is also fixed: $\sigma_k^2 = \tilde{\sigma}^2, \forall k$, where $\tilde{\sigma}^2$ is given by Eq. (3). Thus, the only free parameters in N3 are the relative class frequencies $\pi_k, k = 1, \ldots, K$ and the bias field parameters $\boldsymbol{c}$. We start by analyzing the update equations for $\boldsymbol{c}$.

For the specific scenario where $\sigma_k^2 = \tilde{\sigma}^2, \forall k$, the EM bias field update equation (Eq. (13)) simplifies to

$$\tilde{\boldsymbol{c}} \leftarrow \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi} + 2\tilde{\sigma}^2 \lambda \boldsymbol{\Psi}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{r}, \quad \text{with} \quad \tilde{d}_i = \sum_k w_k^i \tilde{\mu}_k, \quad \boldsymbol{r} = \boldsymbol{d} - \tilde{\boldsymbol{d}},$$

where w_k^i is given by Eq. (9). When the hyperparameter λ is set to the value $\lambda = N\beta/2/\tilde{\sigma}^2$ this corresponds directly to the N3 bias field update equation Eq. (6), where the only difference is that N3 explicitly computes $\tilde{d}_{\mu_l}$ for just 200 discrete intensity values and interpolates to obtain $\tilde{d}_i$, instead of computing $\tilde{d}_i$ directly for each individual voxel.

For the remaining parameters $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^T$, N3 implicitly uses a regularized least-squares fit of the resulting mixture model to the zero-padded normalized histogram $\hat{\boldsymbol{v}}$:

$$\hat{\boldsymbol{\pi}} \leftarrow \underset{\boldsymbol{x}}{\operatorname{argmax}} \|\hat{\boldsymbol{v}} - \boldsymbol{A}\boldsymbol{x}\|^2 + \gamma\|\boldsymbol{x}\|^2, \tag{14}$$

where $\boldsymbol{A}$ is a 512×512 matrix in which each column contains the same Gaussian-shaped basis function, shifted by an offset identical to the column index:

$$\boldsymbol{A} = \begin{pmatrix} g_1 & g_{512} & \cdots & g_2 \\ g_2 & g_1 & \cdots & g_3 \\ \vdots & \vdots & \ddots & \vdots \\ g_{512} & g_{511} & \cdots & g_1 \end{pmatrix},$$

i.e., the first column contains the vector $\boldsymbol{g}$ defined in Eq. (4), and the remaining columns contain cyclic permutations of $\boldsymbol{g}$. To see why Eq. (14) is equivalent to Eq. (2), consider that because $\boldsymbol{A}$ is a circulant matrix, it can be decomposed as

$$\boldsymbol{A} = \boldsymbol{F}^{-1}\boldsymbol{\Lambda}\boldsymbol{F} \quad \text{with} \quad \boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{f}),$$

where $\boldsymbol{F}$ and $\boldsymbol{f}$ were defined in Section 2.1. The solution of Eq. (14) is given by

$$\begin{aligned} \hat{\boldsymbol{\pi}} &\leftarrow \left(\boldsymbol{A}^T\boldsymbol{A} + \gamma\boldsymbol{I}\right)^{-1}\boldsymbol{A}^T\hat{\boldsymbol{v}} = \left(\boldsymbol{F}^{-1}\boldsymbol{\Lambda}^H\boldsymbol{F}\boldsymbol{F}^{-1}\boldsymbol{\Lambda}\boldsymbol{F} + \gamma\boldsymbol{I}\right)^{-1}\boldsymbol{F}^{-1}\boldsymbol{\Lambda}^H\boldsymbol{F}\hat{\boldsymbol{v}} \\ &= \left(\boldsymbol{F}^{-1}\boldsymbol{\Lambda}^H\boldsymbol{\Lambda}\boldsymbol{F} + \gamma\boldsymbol{F}^{-1}\boldsymbol{F}\right)^{-1}\boldsymbol{F}^{-1}\boldsymbol{\Lambda}^H\boldsymbol{F}\hat{\boldsymbol{v}} = \boldsymbol{F}^{-1}\underbrace{\left(\boldsymbol{\Lambda}^H\boldsymbol{\Lambda} + \gamma\boldsymbol{I}\right)^{-1}\boldsymbol{\Lambda}^H}_{\boldsymbol{D}}\boldsymbol{F}\hat{\boldsymbol{v}}, \end{aligned}$$

where $\boldsymbol{A}^H$ denotes the *Hermitian transpose* of $\boldsymbol{A}$ and where we have used the properties that $\boldsymbol{A}^T = \boldsymbol{A}^H$ and $\boldsymbol{F}^H = 512 \cdot \boldsymbol{F}^{-1}$.

An example of N3's mixture model fitted this way will be shown in Figure 1. The periodic end conditions in $\boldsymbol{A}$ have no practical impact on the histogram fit, as the support of the Gaussian-shaped basis functions is limited, and only the parameters of the 200 central basis functions are retained after fitting. Although this is clearly an *ad hoc* approach, the results are certainly not unreasonable, and N3 thereby maintains a close similarity to purely EM-based bias field correction methods.

3 Experiments

Implementation: In order to experimentally verify our theoretical analysis and quantify the effect of replacing the N3 algorithm of Section 2.1 with the EM

algorithm described in Section 2.2 and *vice versa*, we implemented both methods in Matlab. For our implementation of N3, we took care to mimic the original N3 implementation (a Perl script binding together a number of C++ binaries) as faithfully as possible. Specifically, we used identically placed cubic B-spline basis functions $\boldsymbol{\Phi}$, identical regularizer $\boldsymbol{\Psi}$, and the same sub-sampling scheme and parameter settings as in the original method. Our EM implementation shares the same characteristics and preprocessing steps where possible, so that any experimental difference in performance between the two methods is explained by algorithmic rather than technological aspects.

During the course of our experiments, we observed that N3's final basis function fitting operation in the original intensity domain (described in Section 2.1, "Post-processing") actually hurts the performance of the bias field correction. Also, we noticed that N3's default threshold value to detect convergence ($\varsigma = 10^{-3}$) tends to stop the iterations prematurely. To ensure a fair comparison with the EM method, we henceforth report the performance of N3 (Matlab) with the final fitting operation switched off, and with a more conservative threshold value that guarantees full convergence of the method ($\varsigma = 10^{-5}$).

For our EM implementation, we report results for mixture models of $K = 3$, $K = 6$, and $K = 9$ components. We initialize the algorithm with the bias field coefficients set to zero: $\boldsymbol{c} = \mathbf{0}$ (no bias field); with equal relative class frequencies: $\pi_k = 1/K, \forall k$; equidistantly placed means given by Eq. (1) and equal variances given by $\sigma_k^2 = ((\max(\boldsymbol{d}) - \min(\boldsymbol{d}))/K)^2, \forall k$. For a given bias field estimate, the algorithm alternates between re-computing $w_k^i, \forall i, k$ (Eq. (9)) and updating the mixture model parameters (Eqns. (11) and (12)), until convergence in the objective function is detected (relative change between iterations $< 10^{-6}$). Subsequently, the bias field is updated (Eq. 13) and the whole process is repeated until global convergence is detected (relative change in the objective function $< 10^{-5}$).

MRI data and brain masking: We tested both bias field correction methods on two separate datasets of T1-weighted brain MR scans. The first dataset was acquired on several 3T Siemens Tim Trio scanners using a multi-echo MPRAGE sequence with a voxel size of $1.2 \times 1.2 \times 1.2$ mm^3. It consists of 38 subjects scanned twice with varying intervals for a total of 76 volumes. The second dataset consists of 17 volumes acquired on a 7T Siemens whole-body MRI scanner using a multi-echo MPRAGE sequence with a voxel size of $0.75 \times 0.75 \times 0.75$ mm^3. Since N3 bias field correction of brain images is known to work well only on scans in which all non-brain tissue has been removed [11], both datasets were skull-stripped using FreeSurfer[4].

Evaluation metrics: Since the true bias field effect in our MR images is unknown, we compare the two methods using a segmentation-based approach. In particular, we use the coefficient of joint variation [12] in the white and gray matter as an evaluation metric, measured in the original (rather than logarithmic)

[4] `https://surfer.nmr.mgh.harvard.edu/`

domain of image intensities, after bias field correction. This metric is defined as $\text{CJV} = \frac{\sigma_1+\sigma_2}{|\mu_1-\mu_2|}$, where (μ_1, σ_1) and (μ_2, σ_2) denote the mean and standard deviation of intensities within the white and the gray matter, respectively. Compared to the coefficient of variation defined as $\text{CV} = \sigma_1/\mu_1$, which is also commonly used in the literature [11, 13] and which measures only the intensity variation within the white matter, the CJV additionally takes into account the remaining separation between white and gray matter intensities.

In order to compute the CJV, we used FreeSurfer to obtain automatic white and gray matter segmentations, which we then eroded once in order to limit the influence of boundary voxels, which are typically affected by partial volume effects. We observed that the segmentation performance of FreeSurfer was suboptimal in the 7T data because this software has problems with field strengths above 3T. This problem was ameliorated by bias field correcting the 7T scans with SPM8[5] prior to feeding them to FreeSurfer.

In addition to reporting CJV results for the two methods, we also report their run time on a 64bit CentOS 6.5 Linux PC with 24 gigabytes of RAM, an Intel(R) Xeon(R) E5430 2.66GHz CPU, and with Matlab version R2013b installed. For the sake of completeness, we also include the CJV and run time results for the original N3 software (default parameters, with the exception of the spacing between the B-spline control points – see below).

Stiffness of the bias field model: The stiffness of the B-spline bias field model is determined both by the spacing between the B-spline control points (affecting the number of basis functions in $\boldsymbol{\Phi}$) and the regularization parameter of $\boldsymbol{\Psi}$ that penalizes curvature (β in N3, and λ in the EM method).

As recommended in [13], we used a spacing of 50 mm instead of the N3 default[6], as it is known to be too large for images obtained at higher-field strengths. Finding a common, matching value for the regularization parameter in both methods proved difficult, since we observed that the methods perform best in different ranges. Therefore, for the current study we computed average CJV scores for both methods over a wide range of values. We report results for the setting that worked best for each method and for each dataset separately[7].

4 Results

Figure 1 shows the histogram fit and the bias field estimate of both our N3 implementation and the EM method with $K = 6$ Gaussian components on a representative scan from the 7T dataset. In general, the histogram fit works well for both methods; however for N3 a model mismatch can be seen around the high-intensity tail. This is the result of zeroing negative weights after Wiener filtering.

[5] `http://www.fil.ion.ucl.ac.uk/spm/`

[6] 200 mm, appropriate for the 1.5T data the method was originally developed for.

[7] A more elaborate validation study would determine the optimal values on a separate training dataset; however, this is outside the scope of the current workshop paper.

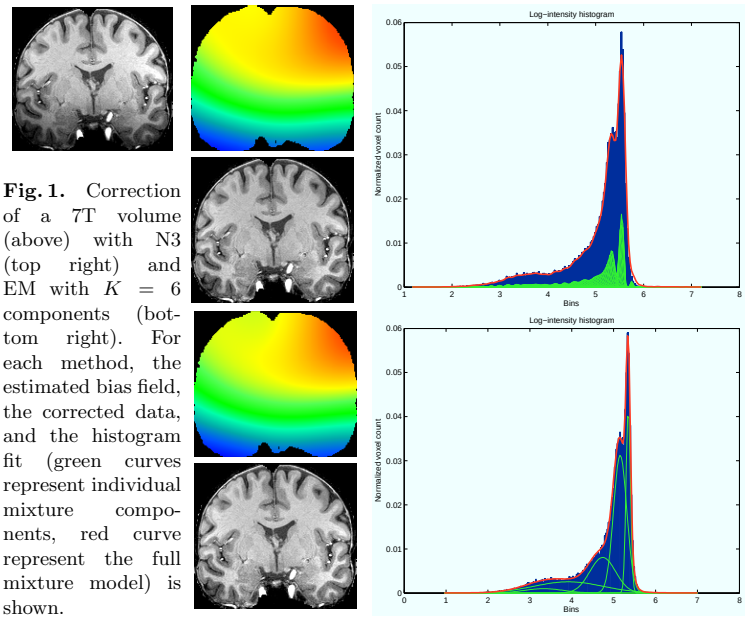


Fig. 1. Correction of a 7T volume (above) with N3 (top right) and EM with $K = 6$ components (bottom right). For each method, the estimated bias field, the corrected data, and the histogram fit (green curves represent individual mixture components, red curve represent the full mixture model) is shown.

Dataset	Average computation time (seconds)				
	EM (3G)	EM (6G)	EM (9G)	N3 (Matlab)	N3
3T	12.7	20.7	29.7	86.0	53.5
7T	50.6	79.2	102.0	415.5	170.8

Table 1. Average computation time for correcting a volume within each dataset.

Figure 2 shows the CJV in the two test datasets, before bias field correction as well as after, using the EM method (for $K = 3$, $K = 6$, and $K = 9$ components), our Matlab N3 implementation, and the original N3 software. Overall, the EM and N3 (Matlab) methods perform comparably, except for EM with $K = 3$ components which seems to have too few degrees of freedom in the 7T dataset. The original N3 implementation is provided as a reference only; its underperformance compared to our own implementation is to be expected since its settings were not tuned the same way.

Table 1 shows the average computation time of each method. Due to the much higher resolution of the 7T data, computation time increased for all methods when correcting this dataset. In all cases, the EM correction ran three to six times faster than the N3 Matlab implementation, depending on the number of components in the mixture. As before, results for the original N3 method are provided for reference only.

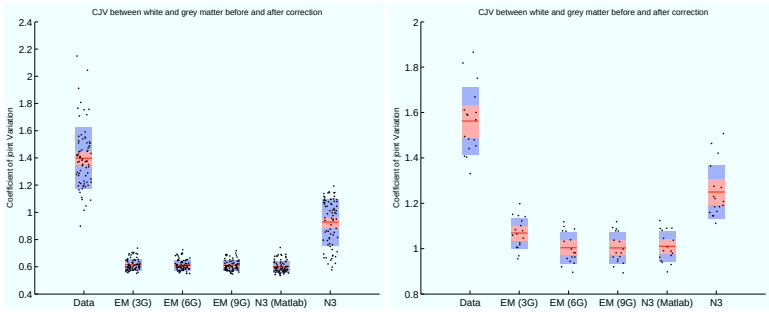


Fig. 2. Scatter plots showing the CVJ between white and gray matter in the 3T (left) and 7T (right) datasets. Lower CVJ equates to better performance. The red line represents the mean, while the blue box covers one standard deviation of the data and the red box covers the 95% confidence interval of the mean.

5 Discussion

In this paper we have explained the successful bias field correction properties of the N3 method by showing that it implicitly uses the same type of generative models and computational strategies as EM-based bias field correction methods. Experiments on MRI scans of healthy brains indicate that, at least in this application, purely EM-based methods can achieve performance similar to N3 at a reduced computational cost.

Future work should evaluate how replacing N3's highly constrained 200-component mixture model with more general mixture models affects bias field correction performance in scans containing pathology. Conversely, while N3's idiosyncratic histogram fitting procedure was found to work well in our experiments, it is worth noting that it precludes N3 from taking advantage of specific prior domain knowledge when such is available. For instance, the skull stripping required to make N3 work well in brain studies [11] typically involves registration of the images into a standard template space, which means that probabilistic brain atlases are available at no additional cost. It is left as further work to evaluate whether this puts N3 at a potential disadvantage compared to EM-based methods, which can easily take this form of extra information into account [3, 7]. Future validation studies should also include comparisons with the publicly available N4ITK implementation [14], which employs a more elaborate but heuristic B-spline fitting procedure in the bias field computations.

Acknowledgments

This research was supported by the NIH NCRR (P41-RR14075), the NIH NIBIB (R01EB013565), TEKES (ComBrain), the Danish Council for Strategic Research (J No. 10-092814) and financial contributions from the Technical University of

Denmark. The authors would like to thank Jonathan Polimeni for supplying 7T data for our tests.

References

1. W. M. Wells, I., Grimson, W.E.L., Kinikis, R., Jolesz, F.A.: Adaptive segmentation of MRI data. IEEE Transactions on Medical Imaging **15**(4) (August 1996) 429 – 442
2. Held, K., Kops, E., Krause, B., Wells, W., Kikinis, R., Muller-Gartner, H.: Markov random field segmentation of brain MR images. IEEE Transactions on Medical Imaging **16**(6) (Dec 1997) 878–886
3. Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P.: Automated model-based bias field correction of MR images of the brain. IEEE Transactions on Medical Imaging **18**(10) (October 1999) 885 – 896
4. Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P.: Automated model-based tissue classification of MR images of the brain. IEEE Transactions on Medical Imaging **18**(10) (October 1999) 897 – 908
5. Pham, D., Prince, J.: Adaptive fuzzy segmentation of magnetic resonance images. IEEE Transactions on Medical Imaging **18**(9) (Sept 1999) 737–752
6. Zhang, Y., Brady, M., Smith, S.: Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. IEEE Transactions on Medical Imaging **20**(1) (2001) 45–57
7. Ashburner, J., Friston, K.J.: Unified segmentation. NeuroImage **26**(3) (July 2005) 839 – 851
8. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) **39**(1) (1977) pp. 1–38
9. Sled, J.G., Zijdenbos, A.P., Evans, A.C.: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Transactions on Medical Imaging **17**(1) (February 1998) 87 – 97
10. Minka, T.P.: Expectation-maximization as lower bound maximization (1998)
11. Boyes, R.G., Gunter, J.L., Frost, C., Janke, A.L., Yeatman, T., Hill, D.L., Bernstein, M.A., Thompson, P.M., Weiner, M.W., Schuff, N., Alexander, G.E., Killiany, R.J., DeCarli, C., Jack, C.R., Fox, N.C.: Intensity non-uniformity correction using N3 on 3-T scanners with multichannel phased array coils. NeuroImage **39**(4) (February 2008) 1752 – 1762
12. Likar, B., Viergever, M.A., Pernus, F.: Retrospective correction of MR intensity inhomogeneity by information minimization. IEEE Transactions on Medical Imaging **20**(12) (Dec 2001) 1398–1410
13. Zheng, W., Chee, M.W., Zagorodnov, V.: Improvement of brain segmentation accuracy by optimizing non-uniformity correction using N3. NeuroImage **48**(1) (2009) 73 – 83
14. Tustison, N., Avants, B., Cook, P., Zheng, Y., Egan, A., Yushkevich, P., Gee, J.: N4ITK: Improved N3 bias correction. IEEE Transactions on Medical Imaging **29**(6) (2010) 1310–1320

Paper C

A unified generative model for MRI bias field correction: re-evaluating and alleviating the need for brainmasking and anatomical atlases

Christian Thode Larsen[a,b], Juan Eugenio Iglesias[c,d], Koen Van Leemput[a,c]

[a]*Department of Applied Mathematics and Computer Science, Technical University of Denmark*
[b]*Danish Research Center for Magnetic Resonance, Denmark*
[c]*Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, MA, USA*
[d]*Basque Center on Cognition, Brain and Language (BCBL), Spain*

Abstract

Correction for intensity inhomogeneity, also referred to as MRI bias field, is often one of the first steps in pipelines for computerized analysis of brain MRI data. Most of the successful methods for bias field correction that are used today depend on probabilistic anatomical atlases, skull stripping or manual user input, or some combination of these in order to achieve good performance. In this paper we present a unified generative model for MRI bias field correction that encompasses many well-known methods as special instances. We experimentally compare the performance of a number of representative methods, using both segmentation-based measures as well as computational speed for evaluation. We also demonstrate the performance of a novel model instance that takes into account how likely voxels are to belong to the same structure given their spatial proximity, thereby alleviating the need for brainmasking or the use of an anatomical atlas.

Keywords: bias field correction, Bayesian inference, generative modeling

1. Introduction

Due to its superior image contrast in soft tissue without involving ionizing radiation, magnetic resonance imaging (MRI) is the *de facto* modality in brain studies, and it is widely used to examine other anatomical regions as well. MRI suffers from an imaging artifact commonly referred to as "intensity inhomogeneity", "intensity bias" or "bias field". The bias field artifact is present at all magnetic field strengths, and is caused by inhomogeneities in B1 transmit field efficiency and receive field sensitivity. While the effects due to the receive field sensitivity depend mostly on the (array of) coils being used for reception, the effects due to the transmit field efficiency increase with field strength (e.g., 7T). The effects due to the transmit field efficiency are dictated by the object being scanned, specifically its shape, position, orientation and the tissue it is be composed of [1, 2, 3]. As a result, while other artifacts which arise from the MRI acquisition device can be corrected using shimming techniques [4, 5], the bias field artifact needs to be estimated for each individual scan using post-processing techniques.

The bias artifact is commonly modeled as a low-frequency, multiplicative effect over the image, an assumption that is only an approximation since inhomogeneities due to the interaction between subject and the transmit field lead to discontinuities in the field over tissue borders. However, the approximation has still proved very useful at field strengths of 1.5 and 3 Tesla, at least for the purpose of segmentation [6].

Bias field correction is a critical step in neuroimaging studies, and is normally performed early in the pipeline, since it avoids the negative impact of intensity inhomogeneities on subsequent computerized analyses. As such, its success is critical for the robustness of the system as a whole, since errors early in the pipeline quickly grow as they propagate through it.

In this study, we propose a generative framework that encompasses a large family of bias field correction algorithms, including many existing methods that can be seen as particular cases of our framework. We also empirically compare the performance of several specific instances, including a novel method that alleviates the need for brainmasking or the use of an anatomical atlas.

1.1. Related work

A broad range of MRI bias field correction methods exist, a detailed overview of which has been pre-

sented in [7]. From a Bayesian point of view, these methods employ models which can be divided (roughly) into three categories. In the first category are generative model-based methods, where the observed image is assumed to be generated from a model given some underlying, unobserved parameters. The parameters of the model are then estimated by maximizing their posterior probability, given the data. In the second category are methods that seek to optimize some heuristic model, such as the frequency content of the histogram. In between these two categories are hybrid methods, which can be shown to employ or relate to an underlying generative model, but where some or all parameters are obtained using heuristic optimization rather than estimating the maximum a-posteriori probability (MAP) parameters. Whereas some methods admittedly have received more interest for the purpose of evaluating bias field correction performance than others, those that have proven to work consistently well all rely on one or several of the following: probabilistic, anatomical atlases, skull stripping or manual user input.

Generative models. These methods commonly integrate bias field correction into tissue classification algorithms, modeling the image intensities using a mixture of Gaussians which are combined with a spatially smooth, multiplicative model of the bias field artifact [8, 9, 10, 11, 12, 13]. Cast as a Bayesian inference problem, fitting these models to the MRI data employs the EM algorithm to estimate some [13] or all [8, 10, 11, 12] of the model parameters. Specifically tailored for brain MRI analysis applications, these methods encode strong prior knowledge about the number and spatial distribution of tissue types present in the images. As such, they yield excellent performance when analyzing brain MRI data, but they cannot be used out of the box to bias field correct images from other anatomical regions.

Heuristic models. Some methods attempt to remove low-frequency components in the image, assumed to be the bias field effect, by means of low-pass filtering techniques [14, 15, 16]. Other methods seek to estimate the field by fitting basis functions to the image data directly, i.e., thin plate splines [17], second-order [18] or fourth-order [19] Legendre polynomials.

Again other methods seek to minimize the entropy of the bias field corrupted image taking into account both multiplicative bias and additive noise [20], or considers only the multiplicative field which is modeled using splines with adaptable control points [21].

A variational level set approach to bias field correction and segmentation is presented in [22], which utilizes a k-means clustering algorithm to partition the data into, and estimate the bias field within, regions in the image domain. More recently, [23] seek to estimate the bias field of 2D MRI images by fitting a Gaussian surface to each of the gradient maps for a number of homogeneous intensity regions, which are selected by automated identification of image histogram peaks.

Hybrid models. In between categories we find N3 [24], arguably the most popular bias field correction method at present. N3 is publicly available and does not use any prior anatomical information on the input, so it can be used for MR scans of any anatomical locations – even if they include pathology. The method is presented as a non-parametric method that maximizes the high-frequency content of the histogram. While this might seem an arbitrary criterion, we have previously shown [25] that N3 is actually parametric, and is based on a generative model which employs a heuristic for parameter estimation. Also receiving increased interest is N4ITK [26], an evolution of N3 that inherits all its advantages, while using a more elaborate and adaptive (yet heuristic) optimization scheme.

[27] presents a fuzzy segmentation scheme that combines tissue classification with bias field correction. The method seeks to minimize an objective function defined as the two-norm between voxel and class intensities weighed with a membership value. Interestingly, this approach is very similar to generative models, as the membership values bears resemblance to the posterior probabilities of class assignments in generative models. Similarly, [28, 29, 30] modify or extend the fuzzy c-means segmentation scheme in order to improve performance, but otherwise preserves the core scheme of the method.

A bias field estimator is formulated by [31] using the conditional probability of observing the data given the bias field effect and a number of global tissue parameters (mean and variance) which are estimated by automated analysis of the image histogram. Based on an assumption of a (small) regionally constant bias field, sample bias field values are then obtained within uniformly positioned regions over the image, by minimizing a cost function on the residual between the observed regional data and corresponding "true" data histograms. These sample values are then smoothed to obtain a global bias field estimate using a regularized least squares fit of cubic b-splines that have been penalized on their bending energy.

1.2. Contribution

The contribution of this paper is threefold. First, we present a unifying generative framework for bias field correction, as well as an associated family of parameter estimation algorithms to compute the bias field based on the generalized expectation-maximization (GEM) algorithm. We further describe instances of this model that correspond to specific bias field correction algorithms, including the "generative" version of N3 that we presented in [25].

Second, we present an extension to the framework that takes advantage of the fact that neighboring voxels typically belong to the same tissue type. Not considering this spatial consistency is a limitation of many current methods, such as N3. The proposed extension is inspired by SLIC superpixels [32] and it aims to substitute or alleviate the absence of probabilistic atlases, which are often necessary to obtain good bias field corrections in cases of severe bias (especially in data at higher field strengths, e.g., 7T). The extension uses an implicit segmentation of the image into supervoxels in the modeling of the inhomogeneities, and it can be easily incorporated to our generative model of bias field correction. Moreover, the extension also makes it unnecessary to mask the region of interest, a step which is typically required to maximize the bias field correction performance [33].

Finally, we provide an extensive empirical evaluation of the different models and corresponding parameter estimation algorithms. We quantitatively compare a total of 12 competing algorithms using longitudinal and cross-sectional MRI data acquired on 3T and 7T scanners.

The rest of this paper is structured as follows. In section 2, we present our unified generative framework. In section 3, we show how the framework can be used to instantiate a number of generative bias field correction models. In section 2.2, we describe relevant parameter updates pertaining to the model instantiations and their optimization using the GEM algorithm. We then present the heuristic parameter optimization scheme to the basic generative model which is used in the popular N3 algorithm.

In section 4 we present a number of experiments that tests the performance of the different models for bias field correction, including speed benchmarks, proxy performance and finally longitudinal performance in Freesurfer. Finally, we discuss the different models, advantages and caveats, results and future work in section 5.

2. General framework for bias field correction

In this section, we present the general framework for bias field correction that we will use throughout the rest of the paper. First, we describe the generative model that we assume for the bias field corrupted MRI data. Then, we propose a GEM algorithm to perform Bayesian inference on the assumed model, in order to obtain an estimate of the bias field.

2.1. Generative model

The generative model is summarized in Figure 1 and Table 1. Let $\boldsymbol{d} = (d_1, \ldots, d_N)^T$ be the log-transformed intensities of the N voxels of a MRI scan of size $N_x \times N_y \times N_z$, and let $\boldsymbol{b} = (b_1, \ldots, b_N)^T$ be the corresponding (log-transformed) gains due to the bias field. Working with log-transformed data is commonplace in the literature, as it simplifies the mathematical analysis by transforming the multiplicative field into an additive one [8, 10, 11, 12]. Therefore, we can write:

$$\boldsymbol{d} = \boldsymbol{u} + \boldsymbol{b},$$

where $\boldsymbol{u} = (u_1, \ldots, u_N)^T$ are the intensities of the "true", uncorrupted, underlying image intensities.

We now assume that $\boldsymbol{b}$ and $\boldsymbol{u}$ are independently generated. Since $\boldsymbol{b}$ varies smoothly in space, we use a linear combination of smooth basis functions, such as cubic B-splines, low order polynomials, or cosine functions. This approximation neglects the fact that the bias field is discontinuous across tissue boundaries, but has been shown to work well [6]. Using this bias field model, we have for M basis functions $\boldsymbol{\phi} = (\phi_{i,1}, \ldots, \phi_{i,M})^T$ evaluated at voxel i and with coefficients $\boldsymbol{c} = (c_1, \ldots, c_M)^T$

$$b_i = \sum_{m=1}^{M} c_m \phi_{i,m}, \tag{1}$$

or, in matrix notation,

$$\boldsymbol{b} = \boldsymbol{\Phi}\boldsymbol{c}, \tag{2}$$

where $\boldsymbol{\Phi}$ is obtained by taking the Kronecker product of matrices composed of M_x, M_y and M_z separable, one-dimensional basis functions evaluated at N_x, N_y and N_z voxels respectively

$$\boldsymbol{\Phi} = \boldsymbol{\Phi}_x \otimes \boldsymbol{\Phi}_y \otimes \boldsymbol{\Phi}_z, \tag{3}$$

with e.g.,

$$\boldsymbol{\Phi}_x = \begin{pmatrix} \phi_{1,1} & \phi_{1,2} & \cdots & \phi_{1,M_x} \\ \phi_{2,1} & \phi_{2,2} & \cdots & \phi_{2,M_x} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{N_x,1} & \phi_{N_x,2} & \cdots & \phi_{N_x,M_x} \end{pmatrix}. \tag{4}$$

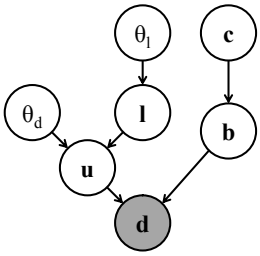


Figure 1: Generative model of bias field corrupted data. Shaded variables are observed.

The bias field coefficients $\boldsymbol{c} = (c_1, \ldots, c_M)^T$ are assumed to be generated by a prior distribution $p(\boldsymbol{c})$.

Regarding the uncorrupted intensities $\boldsymbol{u}$, we assume that they were generated by a mixture of parametric distributions $p(\boldsymbol{u}|\boldsymbol{l}, \boldsymbol{\theta}_d)$ governed by a discrete field of labels $\boldsymbol{l} = (l_1, \ldots, l_N)^T$, which indexes which component generated the intensity at each voxel. The labels l_i, which take discrete values between 1 and L (the total number of labels), cluster together voxels with similar intensity properties, and they might or might not have a direct correspondence with neuroanatomical regions or tissue types. Each label has an associated vector of parameters for the corresponding distribution of intensities (e.g., mean and variance in case of a Gaussian distribution). We will group these parameters into a single vector of parameters $\boldsymbol{\theta}_d$; they are generated by a prior distribution $p(\boldsymbol{\theta}_d)$.

The model is completed by a distribution $p(\boldsymbol{l}|\boldsymbol{\theta}_l)$ of the discrete label field, which depends on parameters $\boldsymbol{\theta}_l$ with their own prior $p(\boldsymbol{\theta}_l)$. The distribution $p(\boldsymbol{l}|\boldsymbol{\theta}_l)$ reflects any prior knowledge on the labels, and could for instance be encoded in a probabilistic atlas, or in a generic smoothness prior. Finally, we assume that both $p(\boldsymbol{l}|\boldsymbol{\theta}_l)$ and $p(\boldsymbol{u}|\boldsymbol{l}, \boldsymbol{\theta}_d)$ factorize over voxels:

$$p(\boldsymbol{l}|\boldsymbol{\theta}_l) = \prod_{i=1}^{N} p(l_i|\boldsymbol{\theta}_l), \quad p(\boldsymbol{u}|\boldsymbol{l}, \boldsymbol{\theta}_d) = \prod_{i=1}^{N} p(u_i|l_i, \boldsymbol{\theta}_d),$$

such that the label l_i and uncorrupted intensity u_i are generated independently for each voxel.

2.2. Inference with GEM

In this section, we describe a GEM algorithm to estimate the parameters in the model above. For convenience, we group all the parameters in a single vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_l^T, \boldsymbol{\theta}_d^T, \boldsymbol{c}^T)^T$. Following Bayes' rule, the maximum a posteriori (MAP) estimate of the parameters is given by:

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\boldsymbol{d}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \left[\log p(\boldsymbol{d}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\right]. \tag{5}$$

$\boldsymbol{\theta}_l$	$\sim$	$p(\boldsymbol{\theta}_l)$
$\boldsymbol{l}$	$\sim$	$p(\boldsymbol{l}\|\boldsymbol{\theta}_l) = \prod_{i=1}^{N} p(l_i\|\boldsymbol{\theta}_l)$
$\boldsymbol{\theta}_d$	$\sim$	$p(\boldsymbol{\theta}_d)$
$\boldsymbol{u}$	$\sim$	$p(\boldsymbol{u}\|\boldsymbol{l}, \boldsymbol{\theta}_d) = \prod_{i=1}^{N} p(u_i\|l_i, \boldsymbol{\theta}_d)$
$\boldsymbol{c}$	$\sim$	$p(\boldsymbol{c})$
$\boldsymbol{b}$	$=$	$\boldsymbol{\Phi c}$
$\boldsymbol{d}$	$=$	$\boldsymbol{u} + \boldsymbol{b}$

Table 1: Generative model of bias field corrupted MRI data.

By exploiting the specific structure of $p(\boldsymbol{d}|\boldsymbol{\theta})$, this optimization can be performed conveniently using a generalized EM (GEM) algorithm [34, 10]. In particular, GEM iteratively builds a lower bound $\varphi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ of the objective function $\log p(\boldsymbol{\theta}|\boldsymbol{d})$ that touches it at the current estimate $\tilde{\boldsymbol{\theta}}$ of the model parameters (E step), and subsequently improves $\varphi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ with respect to the parameters (generalized M step) [34, 35]. This is in contrast with standard EM, in which the bound needs to be exactly optimized at each iteration. In any case, both EM and GEM guarantee that the value of the objective function is increased at each iteration.

Expanding Eq. 5, we obtain:

$$\log p(\boldsymbol{\theta}|\boldsymbol{d}) = \sum_{i=1}^{N} \log\left(\sum_{l=1}^{L} p(d_i|l, \boldsymbol{\theta}_d, \boldsymbol{c}) p(l|\boldsymbol{\theta}_l)\right) \ldots + \log p(\boldsymbol{\theta}_l) + \log p(\boldsymbol{\theta}_d) + \log p(\boldsymbol{c}). \tag{6}$$

Constructing the lower bound of this function (E-step) involves computing soft assignments of each voxel i to each label l (posterior probabilities). The lower bound is given by

$$\varphi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}) = \sum_{i=1}^{N} \sum_{l=1}^{L} w_l^i \log\left(\frac{p(d_i|l, \boldsymbol{\theta}_d, \boldsymbol{c}) p(l|\boldsymbol{\theta}_l)}{w_l^i}\right) \ldots + \log p(\boldsymbol{\theta}_l) + \log p(\boldsymbol{\theta}_d) + \log p(\boldsymbol{c}), \tag{7}$$

with posterior probabilities w_l^i

$$w_l^i = \frac{p(d_i|l, \tilde{\boldsymbol{\theta}}_d, \tilde{\boldsymbol{c}}) p(l|\tilde{\boldsymbol{\theta}}_l)}{\sum_{l'=1}^{L} p(d_i|l', \tilde{\boldsymbol{\theta}}_d, \tilde{\boldsymbol{c}}) p(l'|\tilde{\boldsymbol{\theta}}_l)}. \tag{8}$$

Different GEM algorithms can be obtained by changing the order in which the sets of parameters are optimized in the M step, or the number of times each set of parameters is updated before going back to the E step. Some routes may be more efficient than others, depending on the computational expense of updating each set of parameters.

3. Specific instances

In this section, we present different choices for the probability distributions in the model that yield different bias field correction algorithms. Finally, we present and interpretation of the popular N3 method within our general framework.

3.1. Spatially uniform label priors, Gaussian intensities and Gaussian bias field coefficients

3.1.1. Model

A model that does not have any a priori information on the anatomy in the image to be segmented, can use a spatially uniform prior for the labels, and assign a Gaussian distribution to each of the labels. In that case, the parameter vector $\boldsymbol{\theta}_l$ stores the probabilities π_l for each of the classes, i.e., $\boldsymbol{\theta}_l = (\pi_1, \ldots, \pi_L)^T$, where $\pi_l \geq 0$ and $\sum_l \pi_l = 1$. The prior distribution on the labels is:

$$p(l_i|\boldsymbol{\theta}_l) = \pi_{l_i}.$$

The prior distribution on $\boldsymbol{\theta}_l$ is uniform, i.e., $p(\boldsymbol{\theta}_l) \propto 1$.

For the intensities, the parameter vector $\boldsymbol{\theta}_d$ encompasses the means and variances of L Gaussian distributions, one corresponding to each label: $\boldsymbol{\theta}_d = (\mu_1, \sigma_1^2, \ldots, \mu_L, \sigma_L^2)^T$. The likelihood term becomes:

$$p(d_i|l_i, \boldsymbol{\theta}_d, \boldsymbol{c}) = \mathcal{N}(d_i - b_i|\mu_{l_i}, \sigma_{l_i}^2),$$

where $\mathcal{N}$ is the Gaussian distribution and b_i (which depends on the bias field coefficients $\boldsymbol{c}$) is given by Equation 1. As we do for $\boldsymbol{\theta}_l$, we assume a flat prior distribution for the parameters $\boldsymbol{\theta}_d$, i.e., $p(\boldsymbol{\theta}_l) \propto 1$.

For the bias field coefficients, we use the quadratic prior

$$p(\boldsymbol{c}) \propto \exp[-\lambda \boldsymbol{c}^T \boldsymbol{\Psi} \boldsymbol{c}], \tag{9}$$

where $\boldsymbol{\Psi}$ is a positive semi-definite regularization matrix. Some works (e.g., [10]) use the special case $\lambda = 0$: the smooth nature of the basis functions ensures that the estimated bias field is also smooth. Other works ([8]) use the identity matrix for the basis functions and impose smoothness solely through $\boldsymbol{\Psi}$. However, as shown in Section 4.1.3, further regularization of the field is important when using basis functions with limited support, as the lack of data in the images (e.g., due to the application of a brain mask) may lead to indeterminate equations. Moreover, explicit regularization in the prior can protect against basis functions that are too flexible.

3.1.2. Inference

Setting the partial derivatives of the lower bound with respect to each parameter to zero, it can be shown that the update equations in the M step for this model are:

$$\mu_l \leftarrow \frac{\sum_{i=1}^N w_l^i (d_i - b_i)}{\sum_{i=1}^N w_l^i}, \tag{10}$$

$$\sigma_l^2 \leftarrow \frac{\sum_{i=1}^N w_l^i (d_i - b_i - \mu_l)^2}{\sum_{i=1}^N w_l^i}, \tag{11}$$

$$\pi_l \leftarrow \frac{\sum_{i=1}^N w_l^i}{N}, \tag{12}$$

$$\boldsymbol{c} \leftarrow \left(\boldsymbol{\Phi}^T \boldsymbol{S} \boldsymbol{\Phi} + 2\lambda \boldsymbol{\Psi}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{S} \boldsymbol{r}, \tag{13}$$

where we have defined

$$s_l^i = \frac{w_l^i}{\sigma_l^2}, \quad s_i = \sum_{k=1}^K s_l^i, \quad \boldsymbol{S} = \mathrm{diag}(s_i),$$

$$\bar{d}_i = \frac{\sum_{k=1}^K s_l^i \mu_l}{\sum_{k=1}^K s_l^i}, \quad \boldsymbol{r} = \boldsymbol{d} - \bar{\boldsymbol{d}}.$$

In the special case where all diagonal elements of $\boldsymbol{S}$ are the same $\boldsymbol{S} \propto \boldsymbol{I}$, the update for the bias field coefficients simplifies to:

$$\boldsymbol{c} \leftarrow \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi} + 2\sigma^2 \lambda \boldsymbol{\Psi}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{r}, \tag{14}$$

with

$$\bar{d}_i = \sum_{l=1}^L w_l^i \mu_k, \quad \boldsymbol{r} = \boldsymbol{d} - \bar{\boldsymbol{d}}.$$

3.2. Constrained Gaussian parameters

3.2.1. Model

In cases of severe bias or poor data, better bias field correction may be achieved by constraining the Gaussian parameters in the model in Section 3.1 above, in order to limit the degrees of freedom of the algorithm.

One possibility is to constrain the means of the Gaussian distributions to be equidistant. In that case, the L means are determined by two free parameters μ_1 and μ_L:

$$\mu_l = \alpha_l \mu_1 + (1 - \alpha_l)\mu_L, \tag{15}$$

where $\alpha_l = (l-1)/(L-1)$.

Another option is to force all Gaussians to have the same variance, i.e., $\sigma_l^2 = \bar{\sigma}^2$. The global variance $\bar{\sigma}^2$ can be estimated from the data or set to a fixed predefined value.

3.2.2. Inference

When the means are constrained to be equidistant, we rewrite Eq. 7 as a function of μ_1, μ_L and set derivatives to zero to obtain:

$$\begin{bmatrix}\mu_1\\ \mu_L\end{bmatrix} \leftarrow \begin{bmatrix} \sum_{i=1}^{N}\sum_{l=1}^{L} w_l^i \alpha_l^2 & \sum_{i=1}^{N}\sum_{l=1}^{L} w_l^i (1-\alpha_l)\alpha_l \\ \sum_{i=1}^{N}\sum_{l=1}^{L} w_l^i (1-\alpha_l)\alpha_l & \sum_{i=1}^{N}\sum_{l=1}^{L} w_l^i (1-\alpha_l)^2 \end{bmatrix} \times \begin{bmatrix} \sum_{i=1}^{N}\sum_{l=1}^{L} w_l^i \alpha_l (d_i - b_i) \\ \sum_{i=1}^{N}\sum_{l=1}^{L} w_l^i (1-\alpha_l)(d_i - b_i) \end{bmatrix} \tag{16}$$

The rest of the update equations remain as in Section 3.1.

If we constrain the variances to be equal, it is also necessary to modify the update equation for the variance:

$$\bar{\sigma}^2 \leftarrow \frac{\sum_{i=1}^{N}\sum_{l=1}^{L} w_l^i (d_i - b_i - \mu_l)^2}{N}. \tag{17}$$

3.3. Probabilistic atlas

3.3.1. Model

The model can be augmented with a probabilistic atlas of anatomy describing neuroanatomical structures or tissue types, such as in SPM's "New Segment" [13]. This enables the algorithm to take advantage of this information to potentially produce more accurate corrections, though it also limits its applicability to the brain, or whatever structure the atlas is describing.

When a probabilistic atlas is used, the parameter vector $\boldsymbol{\theta}_l$ stores the atlas probabilities $\{A_{il}\}$, where A_{il} is the a priori probability of observing label l at voxel i according to the atlas. The prior distribution on the labels is then:

$$p(l_i|\boldsymbol{\theta}_l) = A_{il}. \tag{18}$$

Probabilistic atlases typically use few labels that corresponds to the anatomy, e.g., one per tissue type or structure. For this reason, it may be appropriate to use a Gaussian mixture to model the intensities corresponding to each label. Therefore, the vector parameter $\boldsymbol{\theta}_d$ includes, for each label, a predefined number of mixture components K_l, as well as mixture weights π_{lk}, means μ_{lk} and variances σ_{lk}^2 for each component (where l indexes the label and k the mixture component):

$$\begin{aligned}\boldsymbol{\theta}_d = (&\pi_{11}, \ldots, \pi_{1K_1}, \mu_{11}, \ldots, \mu_{1K_1}, \sigma_{11}^2, \ldots, \sigma_{1K_1}^2,\\ &\ldots\\ &\pi_{L1}, \ldots, \pi_{LK_L}, \mu_{L1}, \ldots, \mu_{LK_L}, \sigma_{L1}^2 \ldots, \sigma_{LK_1}^2)^T.\end{aligned}$$

And the likelihood distribution is:

$$p(d_i|l_i, \boldsymbol{\theta}_d, \boldsymbol{c}) = \sum_{k=1}^{K_l} \pi_{lk} \mathcal{N}(d_i - b_i|\mu_{lk}, \sigma_{lk}^2).$$

The model is completed with uniform priors $p(\boldsymbol{\theta}_d) \propto 1$ and $p(\boldsymbol{\theta}_l) \propto 1$.

3.3.2. Inference

As opposed to the instances of our model described in previous sections, the use of a Gaussian mixture for each label requires a slight modification of the E-step (in addition to the M-step) in order to reflect the uncertainty about which mixture component generated the intensity of each voxel given its label. Rather then Eq. 7, the lower bound now is:

$$\begin{aligned}&\varphi(\boldsymbol{\theta}|\boldsymbol{d}) =\\ &\sum_{i=1}^{N}\sum_{l=1}^{L}\sum_{k=1}^{K_l} w_{lk}^i \log\left(\frac{\mathcal{N}(d_i - b_i|\mu_{lk}, \sigma_{lk}^2)\pi_{lk}A_{il}}{w_{lk}^i}\right)\\ &- \lambda \boldsymbol{c}^T \boldsymbol{\Psi} \boldsymbol{c},\end{aligned} \tag{19}$$

where the posteriors now have three indices: voxel, label, and mixture component. In the E-step, these posteriors are computed as:

$$w_{lk}^i = \frac{\mathcal{N}\left(d_i - b_i|\mu_{lk}, \sigma_{lk}^2\right)\pi_{lk}A_{il}}{\sum_{l'=1}^{N}\left[\sum_{k'=1}^{K_l'} \mathcal{N}\left(d_i - b_i|\mu_{l'k'}, \sigma_{l'k'}^2\right)\pi_{l'k'}\right]A_{il'}}. \tag{20}$$

Setting the corresponding derivatives to zero, the parameter updates are given by

$$\mu_{lk} \leftarrow \frac{\sum_{i=1}^{N} w_{lk}^i (d_i - b_i)}{\sum_{i=1}^{N} w_{lk}^i} \tag{21}$$

$$\sigma_{lk}^2 \leftarrow \frac{\sum_{i=1}^{N} w_{lk}^i (d_i - b_i - \mu_{lk})^2}{\sum_{i=1}^{N} w_{lk}^i} \tag{22}$$

$$\pi_{lk} \leftarrow \frac{\sum_{i=1}^{N} w_{lk}^i}{\sum_{i=1}^{N}\sum_{k'=1}^{K_l} w_{lk'}^i}. \tag{23}$$

$$\boldsymbol{c} \leftarrow \left(\boldsymbol{\Phi}^T \boldsymbol{S} \boldsymbol{\Phi} + 2\lambda \boldsymbol{\Psi}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{S} \boldsymbol{r}, \tag{24}$$

with

$$s_{lk}^i = \frac{w_{lk}^i}{\sigma_{lk}^2}, \quad s_i = \sum_{l=1}^{L}\sum_{k=1}^{K_l} s_{lk}^i, \quad \boldsymbol{S} = \mathrm{diag}(s_i),$$

$$\bar{d}_i = \frac{\sum_{l=1}^{L}\sum_{k=1}^{K_l} s_{lk}^i \mu_{lk}}{\sum_{l=1}^{L}\sum_{k=1}^{K_l} s_{lk}^i}, \quad \boldsymbol{r} = \boldsymbol{d} - \bar{\boldsymbol{d}}.$$

3.4. Supervoxels

3.4.1. Model

A way of imposing a generic smoothness constraint in the correction without using a probabilistic atlas, which limits the applicability to a specific anatomical location, is using supervoxels. We propose a model which can be considered a 3-D extension of SLIC Superpixels [32]. Intuitively, the model describes a label as a cluster of voxels of similar intensities within close proximity to each other.

More specifically, we consider a multivariate Gaussian distribution where the covariance matrix is constrained to be non-zero along the diagonal only, such that each dimension is independent. We then define the likelihood of observing both the intensity and spatial position of a voxel given a label similar to [36]:

$$p(\boldsymbol{u}_i|l,\boldsymbol{\theta}_d) = \mathcal{N}\left(\boldsymbol{u}_i|l,\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l\right), \tag{25}$$

with $\boldsymbol{\theta}_d = (\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_L,\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_L)^T$ and exploiting notation, as $\boldsymbol{u}_i = (d_i - b_i, \boldsymbol{x}_i^T)^T$ now is a vector containing the "true" intensity as before as well as the spatial location $\boldsymbol{x}_i$. $\boldsymbol{\mu}_l$ is the mean of the multivariate distribution, and $\boldsymbol{\Sigma}_l = \text{diag}(\sigma^2_{l,intensity}, \sigma^2_{l,x}, \sigma^2_{l,y}, \sigma^2_{l,z})$ is the covariance matrix which encodes the spread of intensities and spatial positions for each supervoxel, separately for each dimension. As in previous configurations, we have $p(l|\boldsymbol{\theta}_l) = \pi_l$ and we assume a uniform prior for the mixture and label parameters $p(\boldsymbol{\theta}_d) \propto 1$ and $p(\boldsymbol{\theta}_l) \propto 1$.

3.4.2. Inference

The inference algorithm is almost identical to that of Section 3.1, except we now also need to update the parameters with respect to the spatial dimensions. Setting the derivatives of Eq.7 to zero yields:

$$\boldsymbol{\mu}_l \leftarrow \frac{\sum_{i=1}^N w_l^i \boldsymbol{u}_i}{\sum_{i=1}^N w_l^i}, \tag{26}$$

where the variance for a single supervoxel along the intensity dimension is

$$\sigma^2_{l,intensity} \leftarrow \frac{\sum_{i=1}^N w_l^i (d_i - b_i - \mu_{l,intensity})^2}{\sum_{i=1}^N w_l^i}. \tag{27}$$

and for each of the spatial dimensions, e.g., for x

$$\sigma^2_{l,x} \leftarrow \frac{\sum_{i=1}^N w_l^i (x_i - \mu_{l,x})^2}{\sum_{i=1}^N w_l^i}. \tag{28}$$

Estimation of the bias field coefficients remain unchanged, as they only depend on the intensities.

3.5. Interpretation of N3 within the proposed framework

Here we describe the popular N3 algorithm as an approximate solver of the model presented in Section 2.

3.5.1. Model

As presented in [25], N3 uses $K = 200$ labels described by Gaussian distributions whose means are equidistantly spaced between the minimum and maximum intensity of the corrected data, i.e., the parameters $\{\mu_l\}$ are fixed. Furthermore, all Gaussians are forced to have an identical variance that is also fixed: $\sigma_k^2 = \bar{\sigma}^2, \forall k$. Therefore, the only free parameters in this model are the relative class frequencies $\pi_k, k = 1,\ldots,K$ and the bias field parameters $\boldsymbol{c}$. This model is very similar to that of Section 3.2.

3.5.2. Approximate Inference

The bias field is computed by alternating between the estimations of $\boldsymbol{\pi}$ and $\boldsymbol{c}$ until convergence. Since the updates are approximate and no objective function is explicitly optimized, there is no guarantee that the algorithm will converge.

Computation of mixture model parameters. The weight parameters are computed by fitting the distribution $p(u|\boldsymbol{\theta}) = \sum_{l=1}^L p_i(u_i|l,\boldsymbol{\theta}_d)p(l|\boldsymbol{\theta}_l)$ to the normalized histogram of the current estimate of the bias field corrected data $\boldsymbol{d} - \tilde{\boldsymbol{d}}$. At each iteration, it is assumed that the centers of the L means are given by:

$$\begin{aligned} &\mu_1 = \min(\boldsymbol{d} - \boldsymbol{b}), \quad \mu_L = \max(\boldsymbol{d} - \boldsymbol{b}), \\ &\mu_l = \mu_1 + (l-1)h, \quad h = (\mu_L - \mu_1)/(L-1), \end{aligned} \tag{29}$$

where h is the bin width. The variance of the Gaussians is fixed and given by:

$$\sigma^2 = \frac{f^2}{8\log 2}, \tag{30}$$

where f denotes a user-specified full-width-at-half-maximum parameter (0.15 by default in N3).

To compute the weights, N3 first computes the histogram entries $\{v_l, l = 1,\ldots,L\}$ using the following interpolation model:

$$v_l = \frac{1}{N}\sum_{i=1}^N \varphi\left[\frac{d_i - b_i - \mu_l}{h}\right], \varphi[s] = \begin{cases} 1-|s| & \text{if } |s|<1 \\ 0, & \text{otherwise.} \end{cases}$$

Then, the weights of the Gaussians are obtained by the following regularized least-squared fit:

$$\boldsymbol{\pi} \leftarrow \underset{\boldsymbol{\pi}}{\text{argmax}} \, \|\hat{\boldsymbol{v}} - \boldsymbol{A}\boldsymbol{\pi}\|^2 + \gamma\|\boldsymbol{\pi}\|^2, \tag{31}$$

where $\hat{\boldsymbol{v}}$ is a padded, 512-dimensional vector such that $\hat{\boldsymbol{v}} = (\mathbf{0}_{156}^T, \boldsymbol{v}^T, \mathbf{0}_{156}^T)^T$, where $\boldsymbol{v} = (v_1, \ldots, v_L)^T$ and $\mathbf{0}_{156}$ is an all-zero 156-dimensional vector. $\boldsymbol{A}$ in Equation 31 is a 512×512 matrix:

$$\boldsymbol{A} = \begin{pmatrix} g_1 & g_{512} & \cdots & g_2 \\ g_2 & g_1 & \cdots & g_3 \\ \vdots & \vdots & \ddots & \vdots \\ g_{512} & g_{511} & \cdots & g_1 \end{pmatrix},$$

in which each column contains the same Gaussian-shaped basis function $\boldsymbol{g}$ shifted by an offset identical to the column index

$$\begin{aligned} \boldsymbol{g} &= (g_1, \ldots, g_{512})^T, \\ g_k &= \begin{cases} h\mathcal{N}((k-1)h|0, \sigma^2) & \text{if } k = 1, \ldots, 256 \\ g_{512-k+1}, & \text{otherwise.} \end{cases} \end{aligned} \quad (32)$$

Vector $\boldsymbol{g}$ is thus a 512-dimensional vector that contains a wrapped Gaussian kernel with variance $\bar{\sigma}^2$.

After $\boldsymbol{\pi}$ has been computed by means of Eq. (31), any negative weights are set to zero, and the padding is removed in order to obtain the central 200-entry weight vector $\boldsymbol{\pi}$. Note that this update is analogous to Eq. 12, though suboptimal in the sense that it does not maximize the bound in Eq. 7.

Computation of bias field coefficients. Given $\boldsymbol{\pi}$, N3 computes expected intensities $\tilde{d}_{\mu_l}$ at each bin center $\tilde{\mu}_l, l = 1, \ldots, L$ as:

$$\tilde{d}_{\mu_l} = \sum_{l'=1}^{L} w_{l'}^{l} \mu_{l'} \quad \text{with} \quad w_{l'}^{l} = \frac{\mathcal{N}\left(\mu_l | \mu_{l'}, \sigma^2\right) \pi_{l'}}{\sum_{l''=1}^{L} \mathcal{N}\left(\mu_l | \mu_{l''}, \sigma^2\right) \pi_{l''}}.$$

Predicted "true" intensities in each voxel are then obtained by linear interpolation between the bin center intensities:

$$\tilde{d}_i = \sum_{l=1}^{L} \tilde{d}_{\mu_l} \varphi\left[\frac{d_i - b_i - \mu_l}{h}\right], \varphi[s] = \begin{cases} 1 - |s| & \text{if } |s| < 1 \\ 0, & \text{otherwise.} \end{cases}$$

These equations are analogous to the E-step of our general framework in Section 2, with the difference that $\tilde{d}_{\mu_l}$ has been computed for just $L = 200$ discrete intensity values and interpolated to obtain $\tilde{d}_i$ – instead of computing $\tilde{d}_i$ directly for each individual voxel.

Finally, a residual $\boldsymbol{r} = \boldsymbol{d} - \tilde{\boldsymbol{d}}$ is computed, and since the variance is equal for all Gaussians $\boldsymbol{S} \propto \boldsymbol{I}$, the bias field coefficients are updated according to Equation 14.

3.6. Convergence aspects

While this heuristic mixture model fitting approach has proven to work well in practice, it does not guarantee an increase in the objective function in every iteration. Where EM algorithms use the objective function to determine convergence, N3 instead uses the standard deviation of the difference in bias estimates between two iterations, and terminate when it drops below a certain threshold (10^{-3} by default).

4. Experiments and results

In this section, we explore the bias field correction performance of a number of configurations of our proposed model, including a version where we employ the heuristic N3 updates for the mixture coefficients.

4.1. Experimental setup

4.1.1. MRI data

We used two different datasets of MRI scans in this study. The first dataset consists of 38 subjects scanned twice with time intervals between two days and six months, for a total of 76 volumes. The data was acquired on several 3T Siemens Tim Trio scanners using identical multi-echo MPRAGE sequences with a voxel size of $1.2 \times 1.2 \times 1.2$ mm^3. The sequences were highly optimized for speed, with a total acquisition time for both scans below five minutes [37]. Time points from the same subject were coregistered, prior to correction, using mri_robust_template from the Freesurfer 5.3 toolbox [38].

The second dataset consists of 30 volumes acquired on a 7T Siemens whole-body MRI scanner equipped with SC72 body gradients using a custom-made 32-channel brain receive coil array and birdcage volume coil for transmit. The data was recorded using a multi-echo MPRAGE sequence with a voxel size of $0.75 \times 0.75 \times 0.75$ mm^3 using the same acquisition parameters as in [39]. Due to the higher field strength, these scans present more severe bias field corruption than the 3T dataset, and represent a bigger challenge for correction methods. This dataset is cross-sectional, so each subject only has a single volume available.

4.1.2. Data preprocessing and parameter initialization

Brainmasks. In order to obtain brain masks for the 3T data, we first preprocessed all volumes using Freesurfer [40]. For the 7T data, the FreeSurfer pipeline does not provide a good mask due to the more intense bias field. To ameliorate this problem, these scans were first bias

field corrected using SPM8[1] and then fed to FreeSurfer for skull stripping.

If bias field correction is performed without skull stripping, it is necessary to filter out background voxels. This is done in order to to avoid estimating the bias field in voxels containing only noise, as these voxels violate the proposed model.

To obtain background-foreground masks, we used the GEM algorithm in Section 3.4 in order to obtain parameter estimates for a number of supervoxels fit to a downsampled version of the image (4mm resolution), using an initial supervoxel grid spacing of 50mm. We then performed Otsu segmentation on the supervoxel intensity means, thereby obtaining the mean intensity threshold value that best separates the supervoxels into background and foreground classes. Each voxel i was finally assigned the label of the supervoxel that it most likely belongs to, i.e., the one that maximizes $\text{argmax}_l\, w_l^i$. As seen in figure 9, this approach effectively removes background voxels, while it preserves voxels containing signal, both brain and non-brain. We found this approach to work well for both the 3T and 7T data.

Parameter initialization. For all models, we initialized the bias field coefficients to zero: $\boldsymbol{c} = \mathbf{0}$ (no bias field). For the configurations where labels were not constrained to equal variance, the Gaussian mixture coefficients were initialized with equal relative frequencies: $\pi_l = 1/L$. In the equal variance configurations, the normalized histogram (using a number of bins equal to the number of labels) were used, as preliminary tests showed slightly better costs at convergence using this approach. The Gaussian means were placed equidistantly between the minimal and maximal intensities of the image, and the variances all set to $\sigma_l^2 = ((\max(\boldsymbol{d}) - \min(\boldsymbol{d}))/L)^2$.

For the configuration using a probabilistic atlas, the model parameters were initialized differently since each label is modeled using separate mixtures. First, we assumed a single Gaussian per class and estimated means and variances, assuming a posterior class probability equal to the corresponding prior probability. Then, the means of the mixture components of each class were initialized by placing them equidistantly in an interval covering three standard deviations from the estimated global mean. The variances were initialized to the squared, equidistant spacing between means, and the mixture coefficients for each label to $\pi_{lk} = 1/K_l$.

In all cases, the model parameters were estimated in a downsampled version of the scans (4mm resolution), and the estimated bias field coefficients subsequently used to compute the bias field at full resolution after convergence.

4.1.3. Smoothness of the bias field

To preserve comparability between all configurations including the one based on the N3 heuristic parameter updates, we used an explicit implementation of cubic B-spline basis functions that mimics the original N3 method exactly, i.e., for the image dimension x:

$$\phi_{x,m} = \sum_{s=0}^{4} \frac{-1^s}{h^3} \binom{4}{s} [x - \lambda_{m-s}]^3 \, \omega\left(x - \lambda_{m-s}\right), \qquad (33)$$

with

$$\omega(k) = \begin{cases} 1, & k \geq 0 \\ 0, & k < 0, \end{cases}$$

where λ_{m-s} is one of $M_x + 3$ knot locations and h is the distance between them. Since cubic B-splines have local support, they are adaptable to local variations in the image.

Because the support of a given b-spline might not cover a sufficient number of foreground voxels to perform the estimation of its coefficient, regularization is necessary. We followed the N3 algorithm and chose a regularization matrix that penalizes the bending energy of the basis functions, generally defined as:

$$J_p(\phi) = \frac{1}{V} \int_{\mathbb{R}^p} \sum_{i=1}^{P} \sum_{j=1}^{P} \left[\frac{\partial^2 \phi}{\partial u_i \partial u_j} \right] d\boldsymbol{u}, \qquad (34)$$

where V is the volume of the region D, expressed in terms of the knot locations:

$$D = \left[\lambda_0^{(x)}, \lambda_{M_x-3}^{(x)}\right] \times \left[\lambda_0^{(y)}, \lambda_{M_y-3}^{(y)}\right] \times \left[\lambda_0^{(z)}, \lambda_{M_z-3}^{(z)}\right]. \qquad (35)$$

Practically, we obtain this matrix by defining:

$$\boldsymbol{\Psi} = \sum_{\substack{\alpha_x,\alpha_y,\alpha_z \geq 0 \\ \alpha_x+\alpha_y+\alpha_z=2}} \frac{2}{\alpha_x!\alpha_y!\alpha_z!} \boldsymbol{\Psi}_x^{(\alpha_x)} \otimes \boldsymbol{\Psi}_y^{(\alpha_y)} \otimes \boldsymbol{\Psi}_z^{(\alpha_z)}, \qquad (36)$$

with elements e.g., for $\boldsymbol{\Psi}_x^{(\alpha_x)}$

$$\psi_{i,j}^{(\alpha)} = \frac{1}{V} \int_D = \phi_i^{(\alpha)}(x) \phi_j^{(\alpha)}(x) \, \mathrm{d}x, \qquad (37)$$

where e.g., $\phi_i^{(\alpha)}$ denotes the α'th derivative of i'th basis function with respect to x. As before, our implementation mimics that of the N3 algorithm exactly.

More generally, smoothness of the bias field is controlled by increasing or decreasing the number of basis

[1] http://www.fil.ion.ucl.ac.uk/spm/

functions M and the regularization hyper parameter λ. Whereas the basis functions only allows a finite number of smoothness levels, the λ hyper parameter in the prior provides continuous control. While it is possible to simply set M to a very high number and then control smoothness using λ, parameter estimation becomes computationally more expensive as M increases. Therefore, one strategy for configuring these two parameters is to first determine the number of basis functions where the model begins to overfit the data, and then increase the regularization in order to impose more smoothness. This is a similar approach to the one used in [33], except they only considered the number of basis functions.

4.1.4. Optimization scheme and assessment of convergence

In the GEM optimization, the algorithm alternates between re-computing the posteriors in the E-step and updating the model parameters in the M-step. During the M-step, the Gaussian parameters are first iteratively updated until the increase in the objective function (Eq. 6) is lower than $< 10^{-6}N$ (here N is the number of voxels). In the case of the N3 updates (Section 3.5), this is a non-iterative process, so no convergence criterion is needed. After the Gaussian parameters have been updated, the bias field coefficients are optimized with Eq. 13. Then, the algorithm begins a new iteration by going back to the E-step. Global convergence is achieved when the standard deviation of the difference bias field between two consecutive iterations is lower than 10^{-5}.

4.1.5. Competing model configurations

A broad range of model configurations were tested:

- **FB-FREE-L3**: Background - foreground segmentation with supervoxel scheme. Mixture composed of $L = 3$ Gaussians with free means, variances and weights, updated according to Eqs. (10), (11) and (12) respectively.
- **FB-FREE-L6**: Same as configuration FB-FREE-L3 above, but with $L = 6$ Gaussians.
- **FB-FIXEDVAR-L20**: Background - foreground segmentation with supervoxel scheme. Mixture composed of $L = 20$ Gaussians with free means and weights, but a single variance updated according to Eq. (17).
- **FB-2MEANS-L200**: Background - foreground segmentation with supervoxel scheme. Mixture of of $L = 200$ Gaussians fitted using equidistant 2-parameter mean update given by equation (16) and fixed variance given equation (30) ($f = 0.15$).

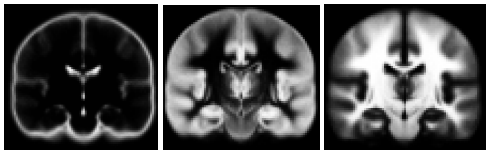


Figure 2: The probabilistic tissue atlas from SPM8. From left to right: CSF, GM and WM.

- **FB-N3-L200**: Background - foreground segmentation with supervoxel scheme.. Mixture of $L = 200$ Gaussians fitted with the N3 updates from Section 3.5 ($f = 0.15$).
- **SUPERVOXELS**: Supervoxel mixture model, using a 50mm grid interval initialization of supervoxel centers. The exact number of Gaussians in the mixture varies, given how many were left after background thresholding. The spatial variance was fixed to the squared spacing between voxel centers along each dimension for all supervoxels.
- **FSM-FREE-L3**: Same as FSM-FREE-L3 above, but using brain the brain extraction from FreeSurfer – rather than the background-foreground segmentation with supervoxels.
- **FSM-FREE-L6**: Same as FB-FREE-L3 above, but with FreeSurfer brain extraction.
- **FSM-FIXEDVAR-L20**: Same as FB-FIXEDVAR-L20 above, but with FreeSurfer brain extraction.
- **FSM-2MEANS-L200**: Same as FB-2MEANS-L200 above, but with FreeSurfer brain extraction.
- **FSM-N3-L200**: Same as FB-N3-L200 above, but with FreeSurfer brain extraction.
- **PROB-ATLAS**: Probabilistic atlas model from Section 3.3. The model uses $L = 4$ labels, which correspond to white matter (2 mixture components), gray matter (2 components), cerebrospinal fluid (2 components) and non-brain tissue (3 components). The label priors are given by the probabilistic tissue atlas from SPM8, illustrated in Figure 2, coregistered to the data using FLIRT version 6,[41, 42, 43]. Note that this model does not require a brain mask; probabilities from the coregistered tissue atlas were used to filter away voxels more than 99% likely to belong non-brain tissue.

4.1.6. Measures of performance

Since the true bias field effect in our MR images is unknown, we use indirect approaches to evaluate the performance of the model configurations: the coefficient of joint variation (which is a segmentation-based approach) and the stability of cortical thickness measures in longitudinal data.

This coefficient of joint variation is defined as CJV = $(\sigma_G + \sigma_W)/(|\mu_G - \mu_W|)$, where (μ_G, σ_W) and (μ_G, σ_W) denote the mean and standard deviation of intensities within the gray and white matter, respectively. Therefore, this metric requires segmentations for the gray and white matter. Compared to the coefficient of variation (defined as CV = σ_W/μ_W), which is also commonly used in the literature [44, 33], the CJV considers not only the intensity variation withing the white matter, but also the separation between the white and gray matter intensities. The CJV was computed in the original (rather than logarithmic) domain of image intensities, after bias field correction.

In the longitudinal data, we use another indirect measure of performance: assuming that the longitudinal scans are close in time, the difference between estimates of cortical thickness (obtained with Freesurfer[40]), can be used to evaluate robustness of the correction method. Following the hypothesis that a better bias field correction will remove more of the bias field, but otherwise not affect the image, the difference in cortical thickness should consequently be closer to zero for the better bias field correction [33].

4.1.7. Experiments

Segmentation based performance with CJV. We bias field corrected all scans in the 3T and 7T datasets using the competing model configurations at two resolution levels of the regularization parameter λ, which describes how much we penalize the flexibility of basis functions when computing the bias field estimate. For all tests, we used a distance setting of 50mm between control points, as presented to be optimal for 3T data in [33]. This may be insufficient on 7T data, and whereas a strategy was previously described for determining the optimal number of basis functions and regularization in Section 4.1.3, we chose this constraint to limit the extent of our testing.

For each dataset, we then progressed to choose the λ values leading to optimal CJV performance using leave-one-out cross validation for the bias field corrected scans. This was done by keeping one scan as a test subject, and the remaining scans in a training set. The mean CJV were computed for all bias field corrected training scans for each value of λ, and the λ leading to the lowest mean CJV across the entire training set were then used to compute and store the corresponding CJV in the test scan. This process was then repeated for each subject in the dataset, and separately for each of the competing configurations, since the optimal value depends of the variance of the mixture model, which in turn depends on the choice of model. The dependence between λ and the variance stems from the fact that higher variances lead to less trust (and therefore lower relative weight) in the fit.

The white and gray matter masks that the CJV requires were automatically obtained with FreeSurfer, and they were eroded with a spherical structuring element of radius 1 in order to limit the influence of boundary voxels on the metric, as these voxels are typically affected by partial volume effects.

Differences in cortical thickness for longitudinal data. We further used the longitudinal 3T data to evaluate the stability of the cortical thickness estimates across time points for the different competing models. The cortical thicknesses were computed with FreeSurfer using the images corrected using the cross-validated, optimal regularization parameters. Comparing the thicknesses of two time points is trivial because FreeSurfer provides a thickness map in standard coordinates.

Computational efficiency. We also recorded the number of iterations that each of the competing algorithms took to converge. Even though directly comparing models with different numbers of parameters is not possible, we can still compare equivalent models with different types of brain masks. Also, we can evaluate the speed of convergence of N3 against that of GEM using the same generative model.

4.2. Results

4.2.1. Segmentation based performance with CJV

Table 2 lists the means and standard deviations of the cross-validated values of λ that lead to the optimal CJV performance for each model configurations, in each of the two datasets. In general, the optimal regularization λ needed to obtain best CJV performance for each model configuration was found to deviate very little between volumes. We observed that more regularization is necessary when masking the brain using FreeSurfer than when segmenting the whole head with supervoxels. This is a consequence of the amount of non-brain voxels in image: the tighter the mask, the less flexibility is required to correct the bias field.

Figure 3 shows the CJVs for the different models for the 3T dataset. We observe that configurations using

		Configurations											
		Foreground-background Mask						Freesurfer Mask					
		FB-FREE-L3	FB-FREE-L6	FB-FIXEDVAR-L20	FB-2MEANS-L200	FB-N3-L200	SUPERVOXELS	FSM-FREE-L3	FSM-FREE-L6	FSM-FIXEDVAR-L20	FSM-2MEANS-L200	FSM-N3-L200	PROB-ATLAS
3T	Reg. (λ)	44.2 (0.4)	32.1 (0.0)	33.6 (0.9)	30.3 (0.7)	32.0 (0.3)	24.2 (0.5)	84.1 (0.2)	80.0 (0.3)	84.1 (0.0)	80.1 (0.4)	47.9 (0.5)	34.1 (0.2)
	Seconds	7 (2)	23 (9)	101 (37)	544 (257)	137 (70)	1112 (281)	2 (0)	10 (6)	47 (12)	272 (35)	47 (7)	14 (4)
	Iterations	21 (10)	56 (24)	176 (100)	289 (175)	246 (125)	542 (220)	19 (5)	24 (5)	93 (23)	108 (11)	151 (20)	26 (5)
7T	Reg. (λ)	2.1 (0.0)	2.1 (0.0)	2.1 (0.0)	2.1 (0.0)	2.1 (0.0)	2.1 (0.0)	44.7 (0.9)	36.0 (0.5)	40.1 (0.0)	34.1 (0.5)	32.0 (0.4)	9.2 (1.0)
	Seconds	13 (4)	55 (16)	237 (37)	2019 (246)	280 (69)	1543 (366)	3 (1)	8 (2)	50 (9)	473 (108)	68 (9)	11 (3)
	Iterations	69 (33)	112 (27)	339 (77)	703 (128)	606 (127)	1419 (495)	31 (7)	35 (6)	114 (40)	245 (40)	251 (30)	33 (5)

Table 2: Mean and standard deviation of the optimal regularization parameter value λ (scaled by a factor 10^{-4} and rounded to one decimal) for each model configuration and dataset, computed with cross validation; and of computational time in seconds as well as iterations necessary to estimate the bias field (both rounded).

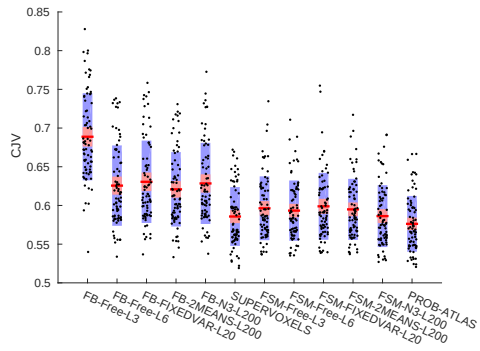


Figure 3: Box plot showing the CVJ between white and gray matter in the 3T dataset. Lower CVJ equates to better performance. The red line represents the mean, while the blue box covers one standard deviation of the data and the red box covers the 95% confidence interval of the mean. The CJV value for each volume in each model were selected by leave-one-out cross validation of the smoothing regularization (λ), leading to the best average CJV for the training set.

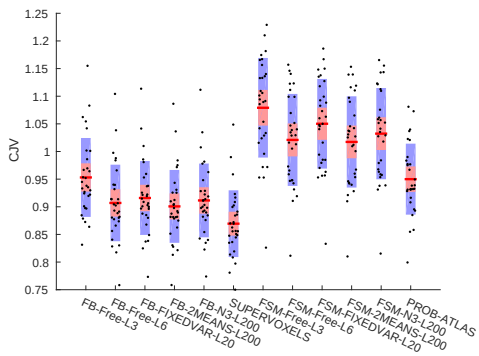


Figure 4: Box plot showing the CVJ between white and gray matter in the 7T dataset. Lower CVJ equates to better performance. The red line represents the mean, while the blue box covers one standard deviation of the data and the red box covers the 95% confidence interval of the mean. The CJV value each volume in each model were found by leave-one-out cross validation of CJV values, given the applied smoothing regularization.

the tight brain mask provided by FreeSurfer outperform the equivalent configurations with the supervoxel-based mask. This finding is in agreement with previous literature on the importance of masking, e.g., [44]. The figure also shows that the configuration with a low number of Gaussian components ($L = 3$) and free weights and variances performs considerably worse than the other configurations. This is due to an insufficient number of degrees of freedom; performance with $L = 6$ is satisfactory. The supervoxel mixture model (see sample in Figure 9) performs comparably to the FreeSurfer masked corrections, despite the fact that it is independent from brain masking. Finally, the configuration using a probabilistic atlas takes advantage of prior knowledge about the image to produce the lower CJV.

On the other hand, the results for the 7T dataset (Figure 4) show superior performance of the supervoxel foreground mask compared over FreeSurfer's brain mask.

Figures 5 and 6 show corrected scans, bias field estimates as well as mixture model fits after one iteration and at convergence for a single 3T and 7T volume for all of the FreeSurfer masked configurations, as well as the supervoxel configuration, all using the

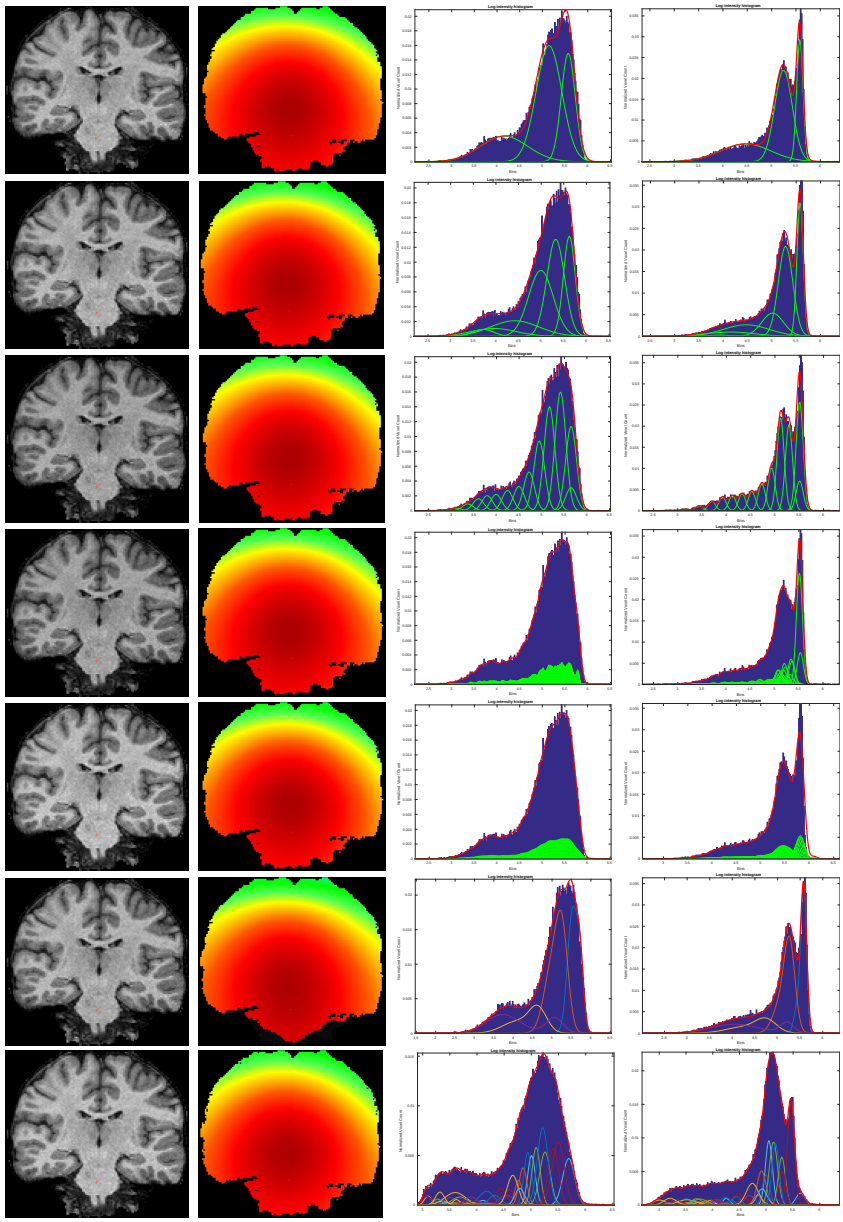


Figure 5: Illustrations of bias field correction using a number of model configurations, given the optimal regularization value, for a single 3T volume. From left to right: corrected data, estimated bias field, mixture fit in the first iteration and at convergence, respectively. Configurations, from top to bottom: FSM-FREE-L3, FSM-FREE-L6 FSM-FIXEDVAR-L20, FSM-2MEANS-L200, FSM-N3-L200, PROB-ATLAS, SUPERVOXELS. Corrected images and associated bias fields all have the Freesurfer mask overlaid, for easy comparability.

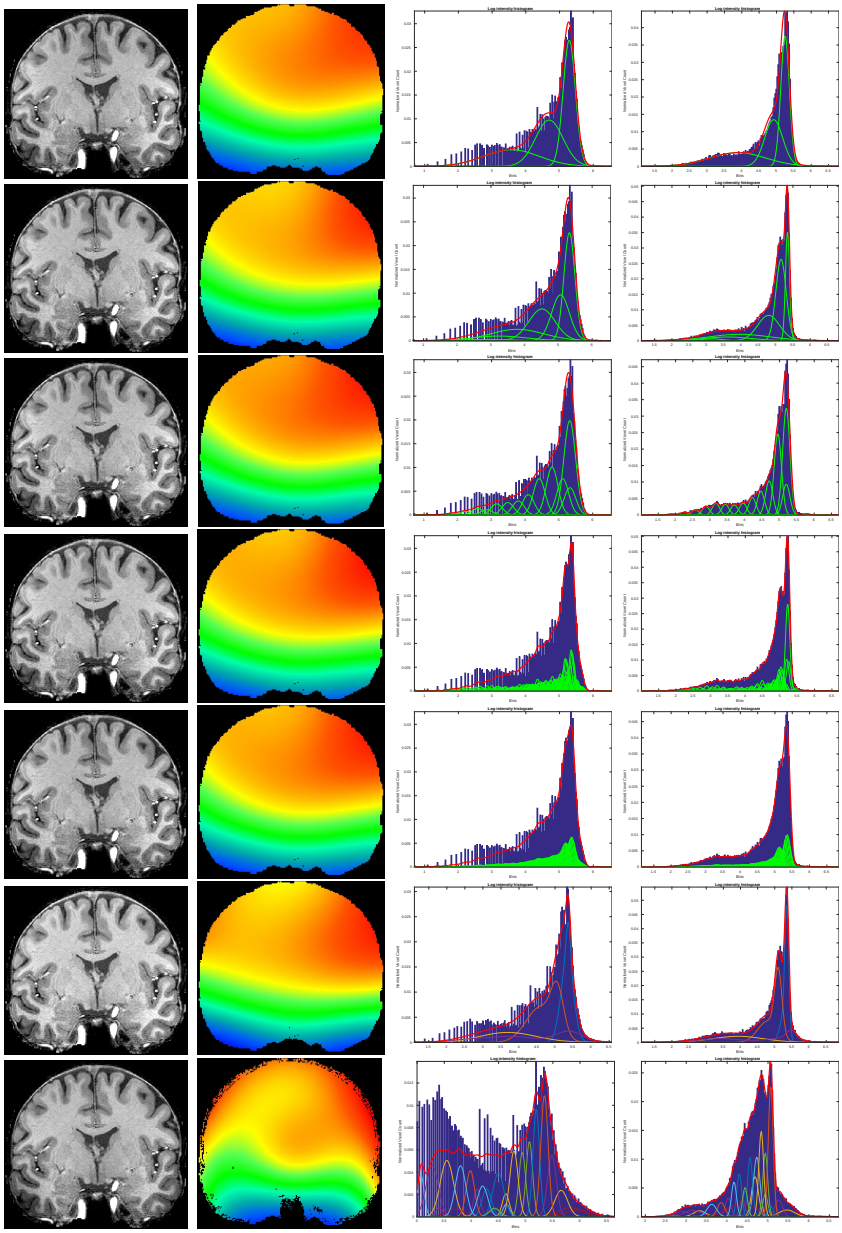


Figure 6: Illustrations of bias field correction using a number of model configurations, given the optimal regularization value, for a single 3T volume. From left to right: corrected data, estimated bias field, mixture fit in the first iteration and at convergence, respectively. Configurations, from top to bottom: FSM-FREE-L3, FSM-FREE-L6 FSM-FIXEDVAR-L20, FSM-2MEANS-L200, FSM-N3-L200, PROB-ATLAS, SUPERVOXELS. Corrected images and associated bias fields all have the Freesurfer mask overlaid, for easy comparability.

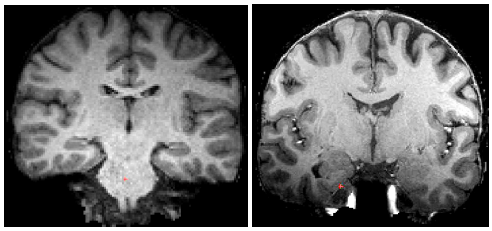


Figure 7: Uncorrected 3T (left) and 7T (right) scans used for the bias field corrections shown in Figure 5 and 6 respectively.

optimal regularization λ. Figure 5 illustrates how the data histogram is highly affected by low intensity voxels in the SUPERVOXELs configuration, but that the configurations otherwise result in very similar fits and corrections. The FSM-N3-L200 configuration appears to have slightly difficulties in fitting the high intensity peak, but this does not seem to impact the quality of the correction. The 7T corrections in Figure 6 show how the FSM-FREE-L3, FSM-FIXEDVAR-L20 and FSM-N3-L200 all have problems fitting the (supposed) gray matter peak. The estimated bias fields are similar for all configurations, except for the SUPERVOXELS configuration. This indicates that the added information in the supervoxel generated mask at 7T results in The uncorrected scans at 3T and 7T corresponding to Figures 5 and 6 have been shown in Figure 7.

4.2.2. Differences in cortical thickness

Figure 8 shows a box plot of the differences in cortical thickness between the two time points per subject for each of the model configurations. It can be seen that the differences are minor ($\sim$ 0.01mm).

4.2.3. Computational efficiency

Table 2 also shows the average computational time as well as number of iterations that it took the GEM algorithm to converge for each competing algorithm. The table reveals that convergence is achieved faster using a tight brain mask, which was expected, since the optimization is only driven by voxels containing relevant information - rather than outliers outside the brain. It is also seen that convergence is achieved faster the more the model is constrained; either by reducing the number of free parameters (which depends on L), or by constraining it using a probabilistic atlas.

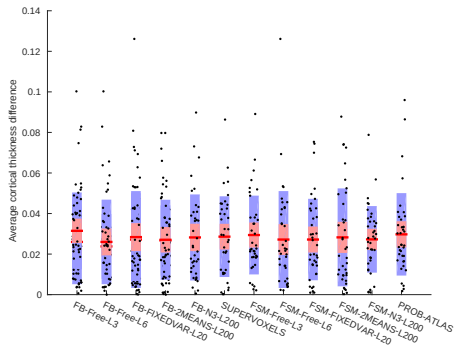


Figure 8: Box plots showing the cortical difference in mm for each model configuration on the 3T dataset. The red line represents the mean, while the blue box covers one standard deviation of the data and the red box covers the 95% confidence interval of the mean.

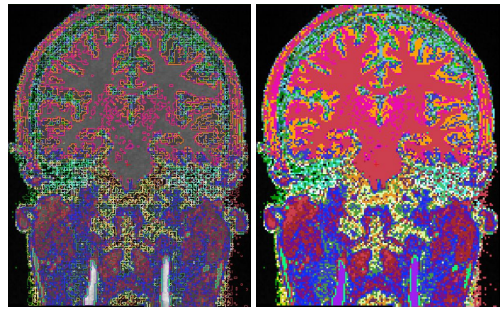


Figure 9: A 3T dataset segmented with the supervoxel mixture model, using an initial grid spacing of 50mm. An outline of the segmented supervoxels is shown to the left, with a filled segmentation to the right.

5. Discussion

Supervoxel model configuration

The supervoxel model extension is highly configurable, given its many parameters and how these can be updated or fixed. Here, we presented results for one configuration where we kept spatial variance fixed, thereby allowing proximity between voxel intensities to dominate the model. A segmentation of the image into the associated supervoxels has been shown in Figure 9. As seen, the configuration produces what appears to be a nice "non-informed" segmentation of the different tissues, which would explain why the model performs as well as it does.

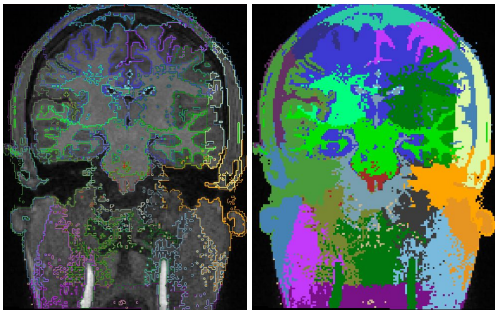


Figure 10: A 3T dataset segmented with the supervoxel mixture model, using an initial grid spacing of 50mm with free variance and fixed mixture coefficients. An outline of the segmented supervoxels is shown to the left, with a filled segmentation to the right.

We also experienced with other configurations, in particular one where we kept the relative label frequencies fixed to equal weight, and then allowed the spatial variance to update. An example has been shown in Figure 10. While this leads to nice image segmentations which bears more resemblance to those shown in [32], preliminary testing of this configuration using the CJV measure suggested that it does not lead to the same performance as when we fix the spatial variance and allow the weights to update. Rather, it performs comparably to the other configurations that depend on background-foreground masking. One possible reason for this is that this configuration is not able to properly take into account the "garbage in the image", i.e., the skull and dura, much similar to the other background-foreground masked configurations.

Bias field correction performance

Whereas the finding that correction of the 7T data is best using the background-foreground mask disagrees with current literature on the importance of masking, the rest of the results suggest that better correction is obtained when the model is properly constrained, in particular by using either a brainmask or a probabilistic tissue atlas.

The supervoxel model configuration shows superior performance in both the 3T and 7T datasets. The fact that the model runs independently of both brainmasking and use of an anatomical atlas, speaks in its favor, in particular in situations where neither is available. However, the model has a very slow convergence rate, in particular at 7T. This is a direct result of the number of free parameters in the model.

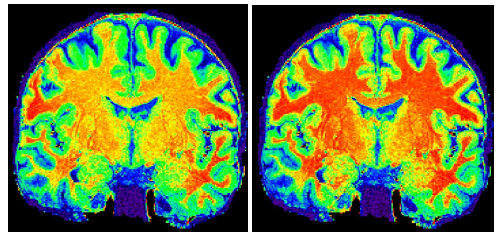


Figure 11: Corrected 7T scan using FSM-FREE-L6 (left) and SUPERVOXELS (right), displayed using a heat map to enhance variations in tissue intensities. It appears as if the supervoxel corrected image is more homogeneous within each tissue, which agrees with the CJV performance.

As a result, the supervoxel should be seen as complementary; if no brainmask or anatomical atlas is available, it provides a nice option. On the other hand, if one or both of the two are available, either can be used in favor of obtaining faster convergence.

Masking at 7T

We performed a thorough inspection of the uncorrected as well as corrected data, and found that the bias field effects in a region around the temporal lobes were particularly severe, leading to very dark voxels. We concluded that these effects result in a misalignment of the bias field when parameters are estimated within the proper brain mask.

Figure 11 shows the data corrected using the FSM-FREE-L6 and SUPERVOXELS configurations. It appears that the tissue is much more homogeneous in the image corrected using the SUPERVOXELS configuration, which agrees with the CJV performance. Further inspection of the corrected data in Figure 6 suggests that the brainmasked corrections may actually be overcompensating for the severe bias field effect, whereas this seem to be alleviated when the model is informed with more data using the supervoxel generated foreground-background masks.

To further verify the validity of the 7T data corrected using the SUPERVOXELS configuration, we also inspected the segmentation into supervoxels as well as the coregistered tissue probability maps. As seen in Figure 12, it appears that the SUPERVOXELS configuration has done a good job of segmenting the image into the respective tissues.

Segmentation performance using Freesurfer

There seem to be no clear connection between the optimal CJV and segmentation performance using

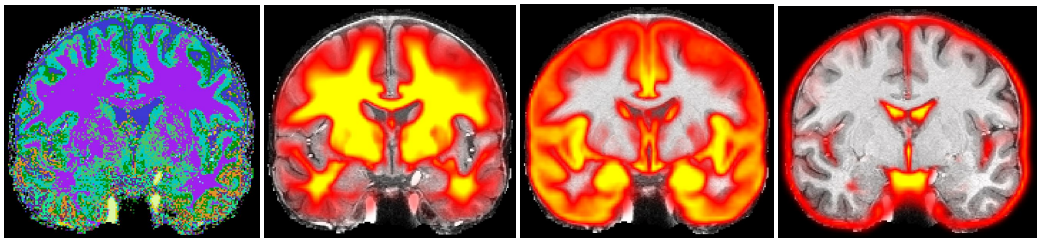


Figure 12: Left to right: Supervoxel segmentation and WM, GM and CSF probability maps respectively coregistered to the 7T scan. The Probability maps appear to be in concensus with the supervoxel segmentation.

Freesurfer. One explanation is that the FS pipeline is so extensive and performs several intensity normalization operations, that it is robust towards relatively small differences in how well GM and WM are separated after bias field correction. As such, it appears that tuning the number of basis functions as well as the regularization in the bias field model is much more important than choosing a particular mixture model – provided that the mixture model fits the data relatively well. Another explanation is any difference in performance is lost because the intervals between cross-validated regularization hyper parameter values were too large.

Future work

It is a logical next step to investigate and validate the software and model underlying the N4ITK algorithm by [26]. They present N4ITK as an evolution of N3, where the underlying cubic B-spline smoothing scheme has been adapted with a more elaborate scheme where control points are allowed to adapt to the image. However, when the generative model behind N3 is considered, the parameter estimates for the bias field coefficients already follows the optimal optimization. This means that the smoothing scheme in N4ITK replaces a valid parameter optimization with a heuristic one, *unless* the more elaborate scheme also can be explained in terms of a generative model.

[26] suggest that N4ITK performs better than N3 given the correlation between bias fields estimated from the Brainweb image generator for a varying number of noise levels and bias field "strengths", and the ground truth. However, the results are not conclusive. First, N3 outperforms N4ITK at the (realistic) noise level of 5% for bias fields that have been scaled in amplitude to field strengths somewhere between 1.5T and 3T.

Second, the default N3 parameters were trained on 1.5T data, *exactly* the field strength where the method outperforms N4ITK on the Brainweb data. Ideally, this training involves cross-validating the optimal distance parameter (number of cubic B-splines) and regularization hyper-parameter. As presented in [25], these parameters, in particular the regularization, need to be retuned at different field strengths and scanners in order to obtain optimal performance, and N3 does not perform optimally at 3T using the default hyper-parameter value. This relationship between the number of basis functions and regularization, and its effect on bias field smoothness, is not considered by [26]. As a result, N3 is, in our opinion, not tested in an optimal way.

Third, the smoothing schemes in the two methods are inherently different, which means you cannot compare the two using the same control point spacing hyper-parameter and expect that performance is comparable. Again, the solution is to employ a cross-validation strategy as suggested.

Finally, the bias fields generated by the Brainweb simulator are not physically correct. While the test setup with respect to the test data is the same for both methods, and therefore can be considered "fair", it remains interesting to compare the methods (including a true generative model implementation) on real MRI data, using e.g., the CJV between WM and GM as the performance measure.

6. Acknowledgements

This work was supported by NIBIB R01EB013565; NIH shared instrumentation grants S10RR023043 and S10RR023401; financial support from the Gipuzkoako Foru Aldundia (Fellows Gipuzkoa Program); the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 654911; the Danish Council for Strategic Research (J No. 10-092814); and financial contributions from the Technical University of Denmark.

References

[1] C. M. Collins, W. Liu, W. Schreiber, Q. X. Yang, M. B. Smith, Central brightening due to constructive interference with, without, and despite dielectric resonance, Journal of Magnetic Resonance Imaging 21 (2) (2005) 192–196.

[2] P.-F. V. de Moortele, C. Akgun, G. Adriany, S. Moeller, J. Ritter, C. M. Collins, M. B. Smith, J. T. Vaughan, K. Uurbil, B1 destructive interferences and spatial phase patterns at 7 t with a head transceiver array coil, Magnetic Resonance in Medicine 54 (6) (2005) 1503–1518.

[3] K. H. Wrede, S. Johst, P. Dammann, L. Umutlu, M. U. Schlamann, I. E. Sandalcioglu, U. Sure, M. E. Ladd, S. Maderwald, Caudal image contrast inversion in {MPRAGE} at 7 tesla: Problem and solution, Academic Radiology 19 (2) (2012) 172 – 178.

[4] Z. Liang, P. C. Lauterbur, Principles of magnetic resonance imaging: a signal processing perspective, IEEE Press series in biomedical engineering, SPIE Optical Engineering Press, Bellingham Wash, 2000, iEEE Engineering in Medicine and Biology Society, sponsor.

[5] Z. Chen, S. S. Li, J. Yang, D. Letizia, J. Shen, Measurement and automatic correction of high-order {B0} inhomogeneity in the rat brain at 11.7 tesla, Magnetic Resonance Imaging 22 (6) (2004) 835 – 842.

[6] M. Styner, K. V. Leemput, Retrospective evaluation and correction of intensity inhomogeneity (2004).

[7] U. Vovk, F. Pernus, B. Likar, A review of methods for correction of intensity inhomogeneity in mri, IEEE Transactions on Medical Imaging 26 (3) (2007) 405–421.

[8] I. W. M. Wells, W. E. L. Grimson, R. Kinikis, F. A. Jolesz, Adaptive segmentation of MRI data, IEEE Transactions on Medical Imaging 15 (4) (1996) 429 – 442.

[9] K. Held, E. Kops, B. Krause, W. Wells, R. Kikinis, H. Muller-Gartner, Markov random field segmentation of brain MR images, IEEE Transactions on Medical Imaging 16 (6) (1997) 878–886.

[10] K. Van Leemput, F. Maes, D. Vandermeulen, P. Suetens, Automated model-based bias field correction of MR images of the brain, IEEE Transactions on Medical Imaging 18 (10) (1999) 885 – 896.

[11] K. Van Leemput, F. Maes, D. Vandermeulen, P. Suetens, Automated model-based tissue classification of MR images of the brain, IEEE Transactions on Medical Imaging 18 (10) (1999) 897 – 908.

[12] Y. Zhang, M. Brady, S. Smith, Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm, IEEE Transactions on Medical Imaging 20 (1) (2001) 45–57.

[13] J. Ashburner, K. J. Friston, Unified segmentation, NeuroImage 26 (3) (2005) 839 – 851.

[14] L. Axel, J. Costantini, J. Listerud, Intensity correction in surface-coil mr imaging, American Journal of Roentgenology 148 (2) (1987) 418–420.

[15] B. Brinkmann, A. Manduca, R. A. Robb, Optimized homomorphic unsharp masking for mr grayscale inhomogeneity correction, Medical Imaging, IEEE Transactions on 17 (2) (1998) 161–171.

[16] M. S. Cohen, R. M. DuBois, M. M. Zeineh, Rapid and effective correction of rf inhomogeneity for high field magnetic resonance imaging, Human Brain Mapping 10 (4) (2000) 204–211.

[17] B. Dawant, A. Zijdenbos, R. Margolin, Correction of intensity variations in mr images for computer-aided tissue classification, Medical Imaging, IEEE Transactions on 12 (4) (1993) 770–781.

[18] C. Brechbühler, G. Gerig, G. Székely, Compensation of spatial inhomogeneity in mri based on a parametric bias estimate, in: K. Hhne, R. Kikinis (Eds.), Visualization in Biomedical Computing, Vol. 1131 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 1996, pp. 141–146.

[19] C. R. Meyer, P. H. Bland, J. Pipe, Retrospective correction of intensity inhomogeneities in mri, Medical Imaging, IEEE Transactions on 14 (1) (1995) 36–41.

[20] B. Likar, M. A. Viergever, F. Pernus, Retrospective correction of MR intensity inhomogeneity by information minimization, IEEE Transactions on Medical Imaging 20 (12) (2001) 1398–1410.

[21] J.-F. Mangin, Entropy minimization for automatic correction of intensity nonuniformity, in: Mathematical Methods in Biomedical Image Analysis, 2000. Proceedings. IEEE Workshop on, IEEE, 2000, pp. 162–169.

[22] C. Li, C. Xu, A. W. Anderson, J. C. Gore, Mri tissue classification and bias field estimation based on coherent local intensity clustering: A unified energy minimization framework, in: Information Processing in Medical Imaging, Springer, 2009, pp. 288–299.

[23] S. Adhikari, J. K. Sing, D. K. Basu, M. Nasipuri, P. Saha, A nonparametric method for intensity inhomogeneity correction in mri brain images by fusion of gaussian surfaces, Signal, Image and Video Processing (2014) 1–10.

[24] J. G. Sled, A. P. Zijdenbos, A. C. Evans, A nonparametric method for automatic correction of intensity nonuniformity in MRI data, IEEE Transactions on Medical Imaging 17 (1) (1998) 87 – 97.

[25] C. T. Larsen, J. Iglesias, K. V. Leemput, N3 bias field correction explained as a bayesian modeling method, in: Bayesian and grAphical Models for Biomedical Imaging, Vol. 8677 of Lecture Notes in Computer Science, Springer International Publishing, 2014, pp. 1–12.

[26] N. Tustison, B. Avants, P. Cook, Y. Zheng, A. Egan, P. Yushkevich, J. Gee, N4ITK: Improved N3 bias correction, IEEE Transactions on Medical Imaging 29 (6) (2010) 1310–1320.

[27] D. Pham, J. Prince, Adaptive fuzzy segmentation of magnetic resonance images, IEEE Transactions on Medical Imaging 18 (9) (1999) 737–752.

[28] M. Ahmed, S. M. Yamany, A. Mohamed, A. A. Farag, T. Moriarty, modified fuzzy c-means algorithm for bias field estimation and segmentation of mri data [j], IEEE Trans. On Medical Imaging 21 (2002) 193–199.

[29] A. W.-C. Liew, H. Yan, An adaptive spatial fuzzy clustering algorithm for 3-d mr image segmentation, Medical Imaging, IEEE Transactions on 22 (9) (2003) 1063–1075.

[30] Z.-X. Ji, Q.-S. Sun, D.-S. Xia, A modified possibilistic fuzzy c-means clustering algorithm for bias field estimation and segmentation of brain mr image, Computerized Medical Imaging and Graphics 35 (5) (2011) 383–397.

[31] D. W. Shattuck, S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, R. M. Leahy, Magnetic resonance image tissue classification using a partial volume model, NeuroImage 13 (5) (2001) 856 – 876.

[32] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Susstrunk, Slic superpixels compared to state-of-the-art superpixel methods, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (11) (2012) 2274 – 2282.

[33] W. Zheng, M. W. Chee, V. Zagorodnov, Improvement of brain segmentation accuracy by optimizing non-uniformity correction using N3, NeuroImage 48 (1) (2009) 73 – 83.

[34] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society. Series B (Methodological) 39 (1) (1977) pp. 1–38.

[35] T. P. Minka, Expectation-maximization as lower bound maxi-

mization (1998).
[36] H. Greenspan, A. Ruf, J. Goldberger, Constrained gaussian mixture model framework for automatic segmentation of mr brain images, IEEE Transactions on Medical Imaging 25 (9) (2006) 1233–1245.
[37] A. J. Holmes, P. H. Lee, M. O. Hollinshead, L. Bakst, J. L. Roffman, J. W. Smoller, R. L. Buckner, Individual differences in amygdala-medial prefrontal anatomy link negative affect, impaired social functioning, and polygenic depression risk, The Journal of Neuroscience 32 (50) (2012) 18087–18100.
[38] M. Reuter, N. J. Schmansky, H. D. Rosas, B. Fischl, Within-subject template estimation for unbiased longitudinal image analysis, NeuroImage 61 (4) (2012) 1402 – 1418.
[39] A. J. van der Kouwe, T. Benner, D. H. Salat, B. Fischl, Brain morphometry with multiecho {MPRAGE}, NeuroImage 40 (2) (2008) 559 – 569.
[40] B. Fischl, Freesurfer, NeuroImage 62 (2) (2012) 774 – 781.
[41] M. Jenkinson, S. Smith, A global optimisation method for robust affine registration of brain images, Medical Image Analysis 5 (2) (2001) 143 – 156.
[42] M. Jenkinson, P. Bannister, M. Brady, S. Smith, Improved optimization for the robust and accurate linear registration and motion correction of brain images, NeuroImage 17 (2) (2002) 825 – 841.
[43] D. N. Greve, B. Fischl, Accurate and robust brain image alignment using boundary-based registration, NeuroImage 48 (1) (2009) 63 – 72.
[44] R. G. Boyes, J. L. Gunter, C. Frost, A. L. Janke, T. Yeatman, D. L. Hill, M. A. Bernstein, P. M. Thompson, M. W. Weiner, N. Schuff, G. E. Alexander, R. J. Killiany, C. DeCarli, C. R. Jack, N. C. Fox, Intensity non-uniformity correction using N3 on 3-T scanners with multichannel phased array coils, NeuroImage 39 (4) (2008) 1752 – 1762.

Bibliography

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. 2012. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence 34*, 11 (Nov), 2274 – 2282.

Adhikari, S., Sing, J. K., Basu, D. K., Nasipuri, M., and Saha, P. 2014. A nonparametric method for intensity inhomogeneity correction in mri brain images by fusion of gaussian surfaces. *Signal, Image and Video Processing*, 1–10.

Ahmed, M., Yamany, S. M., Mohamed, A., Farag, A. A., and Moriarty, T. 2002. modified fuzzy c-means algorithm for bias field estimation and segmentation of mri data [j]. *IEEE Trans. On Medical Imaging 21*, 193–199.

Arnold, J. B., Liow, J.-S., Schaper, K. A., Stern, J. J., Sled, J. G., Shattuck, D. W., Worth, A. J., Cohen, M. S., Leahy, R. M., Mazziotta, J. C., and Rottenberg, D. A. 2001. Qualitative and quantitative evaluation of six algorithms for correcting intensity nonuniformity effects. *NeuroImage 13*, 5, 931 – 943.

Ashburner, J., and Friston, K. J. 2005. Unified segmentation. *NeuroImage 26*, 3 (July), 839 – 851.

Ashburner, J., and Ridgway, G. R. 2013. Symmetric diffeomorphic modeling of longitudinal structural MRI. *Frontiers in NeuroScience 6* (February).

Axel, L., Costantini, J., and Listerud, J. 1987. Intensity correction in surface-coil mr imaging. *American Journal of Roentgenology 148*, 2, 418–420.

Belaroussi, B., Milles, J., Carme, S., Zhu, Y. M., and Benoit-Cattin, H. 2006. Intensity non-uniformity correction in MRI: Existing methods and their validation. *Medical Image Analysis 10*, 2, 234 – 246.

Bookstein, F. L. 2001. "voxel-based morphometry" should not be used with imperfectly registered images. *NeuroImage 14*, 6, 1454 – 1462.

Boyes, R. G., Gunter, J. L., Frost, C., Janke, A. L., Yeatman, T., Hill, D. L., Bernstein, M. A., Thompson, P. M., Weiner, M. W., Schuff, N., Alexander, G. E., Killiany, R. J., DeCarli, C., Jack, C. R., and Fox, N. C. 2008. Intensity non-uniformity correction using N3 on 3-T scanners with multichannel phased array coils. *NeuroImage 39*, 4 (February), 1752 – 1762.

Brechbühler, C., Gerig, G., and Székely, G. 1996. Compensation of spatial inhomogeneity in mri based on a parametric bias estimate. In *Visualization in Biomedical Computing*, K. Hhne and R. Kikinis, Eds., vol. 1131 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 141–146.

Brinkmann, B., Manduca, A., and Robb, R. A. 1998. Optimized homomorphic unsharp masking for mr grayscale inhomogeneity correction. *Medical Imaging, IEEE Transactions on 17*, 2 (April), 161–171.

Burns, A., and Iliffe, S. 2009. Alzheimer's disease. *BMJ 338*.

Castellani, R. J., and Perry, G. 2012. Pathogenesis and disease-modifying therapy in alzheimer's disease: The flat line of progress. *Archives of Medical Research 43*, 8, 694 – 698.

Chen, Z., Li, S. S., Yang, J., Letizia, D., and Shen, J. 2004. Measurement and automatic correction of high-order {B0} inhomogeneity in the rat brain at 11.7 tesla. *Magnetic Resonance Imaging 22*, 6, 835 – 842.

Cohen, M. S., DuBois, R. M., and Zeineh, M. M. 2000. Rapid and effective correction of rf inhomogeneity for high field magnetic resonance imaging. *Human Brain Mapping 10*, 4, 204–211.

Colcombe, S. J., Erickson, K. I., Raz, N., Webb, A. G., Cohen, N. J., McAuley, E., and Kramers, A. F. 2003. Aerobic fitness reduces brain tissue loss in aging humans. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 58*, 2, M176–M180.

Colcombe, S. J., Erickson, K. I., Scalf, P. E., Kim, J. S., Prakash, R., McAuley, E., Elavsky, S., Marquez, D. X., Hu, L., and Kramer, A. F. 2006. Aerobic exercise training increases brain volume in aging humans. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 61*, 11, 1166–1170.

Collins, D., Zijdenbos, A., Kollokian, V., Sled, J., Kabani, N., Holmes, C., and Evans, A. 1998. Design and construction of a realistic digital brain phantom. *IEEE TRANSACTIONS ON MEDICAL IMAGING 17*, 3, 463–468.

Collins, C. M., Liu, W., Schreiber, W., Yang, Q. X., and Smith, M. B. 2005. Central brightening due to constructive interference with, without, and despite dielectric resonance. *Journal of Magnetic Resonance Imaging 21*, 2, 192–196.

Dawant, B., Zijdenbos, A., and Margolin, R. 1993. Correction of intensity variations in mr images for computer-aided tissue classification. *Medical Imaging, IEEE Transactions on 12*, 4 (Dec), 770–781.

de Andrade, L. P., Gobbi, L. T. B., Coelho, F. G. M., Christofoletti, G., Costa, J. L. R., and Stella, F. 2013. Benefits of multimodal exercise intervention for postural control and frontal cognitive functions in individuals with alzheimer's disease: A controlled trial. *Journal of the American Geriatrics Society 61*, 11, 1919–1926.

de Moortele, P.-F. V., Akgun, C., Adriany, G., Moeller, S., Ritter, J., Collins, C. M., Smith, M. B., Vaughan, J. T., and Uurbil, K. 2005. B1 destructive interferences and spatial phase patterns at 7 t with a head transceiver array coil. *Magnetic Resonance in Medicine 54*, 6, 1503–1518.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*, 1, pp. 1–38.

Desikan, R. S., Sgonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., and Killiany, R. J. 2006. An automated labeling system for subdividing the human cerebral cortex on {MRI} scans into gyral based regions of interest. *NeuroImage 31*, 3, 968 – 980.

Destrieux, C., Fischl, B., Dale, A., and Halgren, E. 2010. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage 53*, 1, 1 – 15.

Erickson, K. I., Voss, M. W., Prakash, R. S., Basak, C., Szabo, A., Chaddock, L., Kim, J. S., Heo, S., Alves, H., White, S. M., Wojcicki, T. R., Mailey, E., Vieira, V. J., Martin, S. A., Pence, B. D., Woods, J. A., McAuley, E., and Kramer, A. F. 2011. Exercise training increases size of hippocampus and improves memory. *Proceedings of the National Academy of Sciences 108*, 7, 3017–3022.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., and Dale, A. M. 2002. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron 33*, 3 (January), 341 – 355.

Fischl, B. 2012. Freesurfer. *NeuroImage 62*, 2, 774 – 781.

Fox, N. C., Ridgway, G. R., and Schott, J. M. 2011. Algorithms, atrophy and alzheimer's disease: Cautionary tales for clinical trials. *NeuroImage 57*, 1, 15 – 18.

Fujimoto, K., Polimeni, J. R., van der Kouwe, A. J., Reuter, M., Kober, T., Benner, T., Fischl, B., and Wald, L. L. 2014. Quantitative comparison of cortical surface reconstructions from {MP2RAGE} and multi-echo {MPRAGE} data at 3 and 7 t. *NeuroImage 90*, 60 – 73.

Greenspan, H., Ruf, A., and Goldberger, J. 2006. Constrained gaussian mixture model framework for automatic segmentation of mr brain images. *IEEE Transactions on Medical Imaging 25*, 9 (Sep), 1233–1245.

Gudbjartsson, H., and Patz, S. 1995. The rician distribution of noisy mri data. *Magnetic Resonance in Medicine 34*, 6, 910 – 914.

Guillemaud, R., and Brady, M. 1997. Estimating the bias field of MR images. *IEEE Transactions on Medical Imaging 16*, 3 (June), 238–251.

Hardy, J., and Allsop, D. 1991. Amyloid deposition as the central event in the aetiology of alzheimer's disease. *Trends in Pharmacological Sciences 12*, 0, 383 – 388.

Held, K., Kops, E., Krause, B., Wells, W., Kikinis, R., and Muller-Gartner, H. 1997. Markov random field segmentation of brain MR images. *IEEE Transactions on Medical Imaging 16*, 6 (Dec), 878–886.

Hoffmann, K., Frederiksen, K. S., Sobol, N. A., Nina, B., Vogel, A., Simonsen, A. H., Johannsen, P., Lolk, A., Terkelsen, O., Cotman, C. W., et al. 2013. Preserving cognition, quality of life, physical health and functional ability in alzheimer's disease: the effect of physical exercise (adex trial): rationale and design. *Neuroepidemiology 41*, 3-4, 198–207.

Ji, Z.-X., Sun, Q.-S., and Xia, D.-S. 2011. A modified possibilistic fuzzy c-means clustering algorithm for bias field estimation and segmentation of brain mr image. *Computerized Medical Imaging and Graphics 35*, 5, 383–397.

Kwan, K.-S., Evans, A., and Pike, G. 1999. MRI simulation-based evaluation of image-processing and classification methods. *IEEE Transactions on Medical Imaging 18*, 11, 1085–1097.

Larsen, C. T., Iglesias, J., and Van Leemput, K. 2014. N3 bias field correction explained as a bayesian modeling method. In *Bayesian and grAphical Models for Biomedical Imaging*, vol. 8677 of *Lecture Notes in Computer Science*. Springer International Publishing, 1–12.

Lewis, E. B., and Fox, N. C. 2004. Correction of differential intensity inhomogeneity in longitudinal {MR} images. *NeuroImage 23*, 1, 75 – 83.

Li, C., Xu, C., Anderson, A. W., and Gore, J. C. 2009. Mri tissue classification and bias field estimation based on coherent local intensity clustering: A unified energy minimization framework. In *Information Processing in Medical Imaging*, Springer, 288–299.

Liang, Z., and Lauterbur, P. C. 2000. *Principles of magnetic resonance imaging: a signal processing perspective*. IEEE Press series in biomedical engineering. SPIE Optical Engineering Press, Bellingham Wash. IEEE Engineering in Medicine and Biology Society, sponsor.

Liew, A. W.-C., and Yan, H. 2003. An adaptive spatial fuzzy clustering algorithm for 3-d mr image segmentation. *Medical Imaging, IEEE Transactions on 22*, 9, 1063–1075.

Likar, B., Viergever, M. A., and Pernus, F. 2001. Retrospective correction of MR intensity inhomogeneity by information minimization. *IEEE Transactions on Medical Imaging 20*, 12 (Dec), 1398–1410.

Mangin, J.-F. 2000. Entropy minimization for automatic correction of intensity nonuniformity. In *Mathematical Methods in Biomedical Image Analysis, 2000. Proceedings. IEEE Workshop on*, IEEE, 162–169.

Meyer, C. R., Bland, P. H., and Pipe, J. 1995. Retrospective correction of intensity inhomogeneities in mri. *Medical Imaging, IEEE Transactions on 14*, 1, 36–41.

Minka, T. P., 1998. Expectation-maximization as lower bound maximization.

Modat, M., Ridgway, G. R., Hawkes, D. J., Fox, N. C., and Ourselin, S. 2010. Nonrigid registration with differential bias correction using normalised mutual information. In *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on*, IEEE, 356–359.

Nikolaev, A., McLaughlin, T., OLeary, D. D., and Tessier-Lavigne, M. 2009. App binds dr6 to trigger axon pruning and neuron death via distinct caspases. *Nature 457*, 7232, 981–989.

Pham, D., and Prince, J. 1999. Adaptive fuzzy segmentation of magnetic resonance images. *IEEE Transactions on Medical Imaging 18*, 9 (Sept), 737–752.

Reuter, M., Rosas, H. D., and Fischl, B. 2010. Highly accurate inverse consistent registration: A robust approach. *NeuroImage 53*, 4, 1181 – 1196.

Reuter, M., Schmansky, N. J., Rosas, H. D., and Fischl, B. 2012. Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage 61*, 4 (July), 1402 – 1418.

Ruscheweyh, R., Willemer, C., Krger, K., Duning, T., Warnecke, T., Sommer, J., Vlker, K., Ho, H., Mooren, F., Knecht, S., and Flel, A. 2011. Physical activity and memory functions: An interventional study. *Neurobiology of Aging 32*, 7, 1304 – 1319.

Shackleford, J. A., Yang, Q., Loureno, A. M., Shusharina, N., Kandasamy, N., and Sharp, G. C. 2012. Analytic regularization of uniform cubic b-spline deformation fields. In *Medical Image Computing and Computer-Assisted Intervention MICCAI 2012*, N. Ayache, H. Delingette, P. Golland, and K. Mori, Eds., vol. 7511 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 122–129.

Shattuck, D. W., Sandor-Leahy, S. R., Schaper, K. A., Rottenberg, D. A., and Leahy, R. M. 2001. Magnetic resonance image tissue classification using a partial volume model. *NeuroImage 13*, 5, 856 – 876.

Sled, J. G., Zijdenbos, A. P., and Evans, A. C. 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging 17*, 1 (February), 87 – 97.

Stöcker, T., Vahedipour, K., Pflugfelder, D., and Shah, N. J. 2010. High-performance computing mri simulations. *Magnetic Resonance in Medicine 64*, 1, 186–193.

Styner, M., and Van Leemput, K., 2004. Retrospective evaluation and correction of intensity inhomogeneity.

Thompson, W. K., and Holland, D. 2011. Bias in tensor based morphometry stat-roi measures may result in unrealistic power estimates. *NeuroImage 57*, 1, 1 – 4.

Tustison, N., Avants, B., Cook, P., Zheng, Y., Egan, A., Yushkevich, P., and Gee, J. 2010. N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging 29*, 6, 1310–1320.

Van Leemput, K., and Puonti, O. 2015. Tissue classification. In *Brain Mapping*, A. W. Toga, Ed. Academic Press, Waltham, 373 – 381.

Van Leemput, K., Maes, F., Vandermeulen, D., and Suetens, P. 1999. Automated model-based bias field correction of MR images of the brain. *IEEE Transactions on Medical Imaging 18*, 10 (October), 885 – 896.

Van Leemput, K., Maes, F., Vandermeulen, D., and Suetens, P. 1999. Automated model-based tissue classification of MR images of the brain. *IEEE Transactions on Medical Imaging 18*, 10 (October), 897 – 908.

Van Leemput, K., Bakkour, A., Benner, T., Wiggins, G., Wald, L. L., Augustinack, J., Dickerson, B. C., Golland, P., and Fischl, B. 2009. Automated segmentation of hippocampal subfields from ultra-high resolution in vivo mri. *Hippocampus 19*, 6, 549–557.

Velthuizen, R. P., Heine, J. J., Cantor, A. B., Lin, H., Fletcher, L. M., and Clarke, L. P. 1998. Review and evaluation of MRI nonuniformity corrections for brain tumor response measurements. *Medical physics 25*, 1655.

Vovk, U., Pernus, F., and Likar, B. 2007. A review of methods for correction of intensity inhomogeneity in mri. *IEEE Transactions on Medical Imaging 26*, 3 (March), 405–421.

Vreugdenhil, A., Cannell, J., Davies, A., and Razay, G. 2012. A community-based exercise programme to improve functional ability in people with alzheimers disease: a randomized controlled trial. *Scandinavian Journal of Caring Sciences 26*, 1, 12–19.

W. M. Wells, I., Grimson, W. E. L., Kinikis, R., and Jolesz, F. A. 1996. Adaptive segmentation of MRI data. *IEEE Transactions on Medical Imaging 15*, 4 (August), 429 – 442.

Whitwell, J. L., Przybelski, S. A., Weigand, S. D., Knopman, D. S., Boeve, B. F., Petersen, R. C., and Jack, C. R. 2007. 3d maps from multiple mri illustrate changing atrophy patterns as subjects progress from mild cognitive impairment to alzheimer's disease. *Brain 130*, 7, 1777–1786.

Wrede, K. H., Johst, S., Dammann, P., Umutlu, L., Schlamann, M. U., Sandalcioglu, I. E., Sure, U., Ladd, M. E., and Maderwald, S. 2012. Caudal image contrast inversion in {MPRAGE} at 7 tesla: Problem and solution. *Academic Radiology 19*, 2, 172 – 178.

Zhang, Y., Brady, M., and Smith, S. 2001. Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging 20*, 1, 45–57.

Zheng, W., Chee, M. W., and Zagorodnov, V. 2009. Improvement of brain segmentation accuracy by optimizing non-uniformity correction using N3. *NeuroImage 48*, 1, 73 – 83.